Validation of credit risk models

Review and application of key validation tests

Collaboration between Human and AI

February 7, 2025

Contents

1 Introduction to Credit Risk

Credit risk is the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations. It is a fundamental concern for financial institutions, as it directly impacts their profitability and solvency. Understanding and managing credit risk is essential for maintaining the stability of the financial system.

Financial institutions employ various credit risk models to assess and quantify the risk associated with lending activities. These models help in estimating the probability of default (PD), loss given default (LGD), and exposure at default (EAD), which are critical components in determining capital requirements and risk management strategies.

Risk differentiation is a key aspect of credit risk modeling. It involves identifying relevant risk drivers and using them to rank or differentiate obligors or exposures into grades or pools according to their level of risk. This process enables institutions to tailor their credit decisions and pricing strategies based on assessed risk levels.

Credit risk models can be broadly categorized based on the types of exposures they address:

- **Corporate Models**: Focused on exposures to corporate entities, including both general corporates and specialized lending. Specialized lending refers to financing activities where the repayment depends primarily on the income generated by the assets being financed.
- **Institution Models**: Designed to assess the credit risk associated with exposures to financial institutions, such as banks and insurance companies.
- **Retail and SME Models**: Targeted at portfolios consisting of retail customers and small and medium-sized enterprises (SMEs).

Model validation plays a crucial role in ensuring the accuracy and reliability of credit risk models. Credit institutions are required to develop validation processes to assess and validate their internal models. While all validation processes must comply with the same regulatory requirements, the ability to perform comparisons across models and institutions remains limited. Most validation tools focus on the quantitative aspects of internal models. To provide clarity and facilitate the implementation of these tools, detailed instructions are essential to reduce ambiguity and ensure consistency.

Expressing validation results through ratings allows relevant information to be easily incorporated into decision-making processes and trigger appropriate actions. Effective model validation contributes to identifying and mitigating model risk—the risk of adverse consequences arising from decisions based on incorrect or misused models.

In this part of the book, we will delve into the various credit risk models used for different exposure classes, including corporates (both general and specialized lending) and institutions. Similar to models for retail and SME portfolios, the selection of low default portfolio (LDP) models for discussion is based on an assessment of their materiality and criticality.

Understanding credit risk and the models used to measure it is vital for financial institutions to manage their risk exposures effectively. Moreover, robust model validation processes are essential to ensure the models' reliability, thereby enhancing the overall stability of the financial system.

1.1 Credit Risk Model Types

Credit risk models are essential tools in financial risk management, enabling institutions to quantify and manage potential losses arising from borrowers' defaults. The primary types of credit risk models include the Probability of Default (PD), Loss Given Default (LGD), Exposure at Default (EAD), and Expected Loss Best Estimate (ELBE). Each of these models serves a specific purpose in assessing different dimensions of credit risk.

Probability of Default (PD) models estimate the likelihood that a borrower will default on their obligations within a specified time horizon. These models involve data and methodologies used to assess default risk for each obligor or exposure. PD estimates are integral to credit decision-making processes and regulatory capital calculations, as they help institutions quantify default risk and determine appropriate levels of capital reserves.

Loss Given Default (LGD) models quantify the proportion of an exposure that a lender expects to lose if a borrower defaults, after accounting for recoveries from collateral or other credit enhancements. LGD estimation often involves analyzing various components, such as the secured and unsecured portions of the exposure, termination scenarios, and different phases in the recovery process (e.g., before and after legal proceedings). Accurate LGD models are crucial for determining potential loss severities and are used in calculating expected losses and economic capital.

Exposure at Default (EAD) models estimate the amount outstanding at the time of default. For off-balance-sheet items like credit lines and loan commitments, EAD models predict the extent to which these facilities will be utilized before default occurs. Reliable EAD estimates are vital for calculating expected losses and meeting regulatory capital requirements.

Expected Loss Best Estimate (ELBE) provides an estimation of the expected loss for exposures that have already defaulted. ELBE models assess the anticipated loss by considering factors such as current economic conditions and the specific characteristics of the defaulted exposure. In practice, ELBE estimates may be derived from dedicated models or set equal to specific credit risk adjustments for the exposure. In some cases, institutions base ELBE estimations on LGD models designed for performing exposures or utilize empirical evidence from internal data. ELBE is used in the calculation of provisions and in determining the appropriate level of capital to hold against defaulted assets.

The development and implementation of these credit risk models are critical for effective risk management and regulatory compliance. They enable financial institutions to better understand their credit risk profile, make informed lending decisions, allocate capital efficiently, and comply with regulatory standards. The accuracy and robustness of PD, LGD, EAD, and ELBE models directly impact an institution's ability to manage credit risk and maintain financial stability.

1.2 Model Validation Importance

Model validation is a critical process in ensuring the *accuracy*, *reliability*, and *regulatory compliance* of models used within the financial industry. In various fields such as computer science, engineering, and finance, model validation refers to the key assessments undertaken to verify that a model is working as expected. This is particularly important in finance, where models play a central role in decision-making processes.

Model risk can be described as the potential for adverse consequences arising from decisions based on incorrect or misused model results and reports. According to point 11 of Article 3(1) of the Capital Requirements Directive (Directive 2013/36/EU - CRD), model risk is defined as the risk of a potential loss an institution may incur due to errors in the development, implementation, or use of internal models. The main task of the model validation process is to prevent models from producing inadequate results by effectively challenging them and assessing possible assumptions, limitations, and shortcomings.

In the era of advanced analytics, Machine Learning (ML) models offer significant added value. However, to comply with the Capital Requirements Regulation (CRR) requirements, it is essential that these models are interpretable. Relevant stakeholders must have a level of knowledge of the model's functioning that is at least proportionate to their involvement and responsibility for meeting legal requirements. Otherwise, there is a risk of developing *black box* models, where the internal workings are not transparent, leading to potential misinterpretation and misuse.

A sound validation function is crucial for:

- Ensuring the reliability of internal models.
- Accurately computing capital requirements.
- Preventing potential losses due to model errors.
- Maintaining compliance with regulatory requirements.

It is the responsibility of the credit institution to ensure that its internal models are fully compliant with all regulatory requirements. This includes ongoing model monitoring tools such as benchmarking and the review of validation reports, which help identify areas where further investigations are needed. Rigorous off-site assessments can complement on-site investigations for less material or less critical models or model changes, allowing for a more risk-based approach.

Furthermore, convergence in supervisory methodologies and practices for assessing internal models is key for:

- Ensuring comparability of models' outcomes.
- Restoring public confidence in the use of models for regulatory purposes.
- Establishing consistent evaluation criteria, frequency, and scope of analysis.
- Ensuring a level playing field in the Single Market.

Institutions and all levels of their management functions and bodies must have an adequate understanding of their Internal Ratings-Based (IRB) models. This understanding is essential for allowing ML models to be used for regulatory purposes while avoiding the pitfalls associated with complex, non-transparent models.

In summary, model validation is indispensable for financial institutions to:

- Mitigate model risk.
- Enhance decision-making processes.
- Achieve regulatory compliance.
- Promote transparency and accountability.

By effectively implementing robust model validation processes, institutions can safeguard against potential losses, ensure regulatory compliance, and contribute to the overall stability of the financial system.

1.3 Model Risk Implications

Model risk poses significant threats to financial institutions, encompassing potential financial losses, reputational damage, and regulatory penalties. Following the financial crisis of 2007–09, concerns intensified regarding the unwarranted variability of outputs from models used to calculate regulatory capital requirements. This variability, not always based on actual risk differences, raised alarms about the complexity and opaqueness of modeling approaches, making it challenging for supervisors to assess risk capture accurately and consistently.

Effective model risk management is crucial in mitigating these threats. Institutions should implement a comprehensive model risk management framework that enables them to identify, understand, and manage model risk across all models used internally. Such a framework should include:

- **Identification and Mitigation of Model Deficiencies**: Establish guidelines for identifying areas where measurement uncertainty and model deficiencies exist, prioritizing them according to materiality. This includes addressing qualitative aspects of model risk, such as data deficiencies, model misuse, or implementation errors.
- **Assessment and Measurement of Model Risk**: Develop methodologies for both qualitative and quantitative assessment of model risk. This enables institutions to gauge the potential impact of model inaccuracies on their financial standing and regulatory compliance.

Failing to manage model risk effectively can lead to:

- **Financial Losses**: Inaccurate models may lead to mispricing of assets, inadequate provisioning, or misallocation of capital, resulting in direct financial losses.

- **Reputational Damage**: Stakeholders may lose confidence in an institution's risk management capabilities, adversely affecting the institution's market position and stakeholder relationships.
- **Regulatory Penalties**: Non-compliance with regulatory requirements, such as those outlined in the Capital Requirements Regulation (CRR), can result in penalties, increased capital requirements, or other supervisory actions.

By focusing on robust model risk management practices, institutions can reduce the likelihood of these adverse outcomes. This involves not only meeting prudential requirements but also considering ethical and legal aspects, as well as consumer and data protection, although those areas are beyond the scope of this discussion. Prioritizing transparency and simplicity in modeling approaches can further enhance the ability of supervisors to assess risk capture effectively, thereby fostering greater confidence among external stakeholders.

1.4 Model Lifecycle

The model lifecycle encompasses the comprehensive process through which a financial model is developed, implemented, validated, utilized, monitored, and eventually retired or replaced. Understanding each stage is crucial to ensure the model's effectiveness, compliance with regulatory standards, and alignment with the institution's objectives.

1. **Development**: This initial phase involves defining the model's purpose, scope, and design. Critical activities include data preparation—such as data collection, cleaning, and transformation—and selecting appropriate methodologies. The goal is to build a robust model that accurately represents the financial phenomena it aims to capture.
2. **Implementation**: After development, the model is integrated into the institution's internal processes. This step may require supervisory approval if necessary. Implementation ensures that the model operates effectively within existing systems and that users are adequately trained to apply it correctly.
3. **Validation**: Independent validation is conducted to assess the model's performance and reliability. Validators examine the model's assumptions, testing procedures, and outcomes to confirm that it meets all regulatory requirements and internal standards. Validation helps identify any weaknesses that need to be addressed before full deployment.
4. **Use**: The validated model is deployed for its intended purposes, such as risk assessment, credit evaluation, or regulatory capital calculation. Proper application of the model supports decision-making processes and enhances the institution's risk management capabilities.
5. **Monitoring**: Continuous monitoring is essential to ensure the model remains accurate over time. This involves regularly reviewing model outputs, performance metrics, and underlying data. Monitoring activities help detect any degradation in model performance due to changes in market conditions or internal processes.

6. **Retirement or Replacement**: When a model no longer provides reliable results or becomes obsolete due to new technologies or methodologies, it must be retired or replaced. This process should be carefully managed to transition to a new model without disrupting operations. Retirement decisions are based on comprehensive reviews and align with regulatory guidelines.

Throughout the model lifecycle, maintaining up-to-date documentation is imperative. Institutions should keep detailed records of all stages, including development decisions, validation results, and monitoring activities. Documentation must be retained for appropriate periods in compliance with legal or regulatory retention requirements, ensuring transparency and facilitating audits or supervisory evaluations.

1.5 Key Terminology

In this section, we define key terms essential for understanding credit risk modelling and validation, incorporating definitions from the European Central Bank (ECB) regulatory documents and general statistical concepts.

- **Credit Risk**: The risk of loss resulting from a borrower's inability or unwillingness to repay a loan or meet contractual obligations.
- **Probability of Default (PD)**: The likelihood that a borrower will default on their obligations within a specified time horizon, typically one year.
- **Loss Given Default (LGD)**: The proportion of an exposure that is expected to be lost if a borrower defaults. It represents the severity of loss when default occurs.
- **Exposure at Default (EAD)**: The total value a bank is exposed to at the time of a borrower's default.
- **Expected Loss Best Estimate (ELBE)**: An estimate of the expected loss on defaulted assets, reflecting the actual economic loss anticipated.
- **LGD for Defaulted Assets (LGD in-default)**: The LGD estimation specifically for assets that have already defaulted, used in calculating capital requirements for defaulted exposures.
- **Credit Conversion Factor (CCF)**: A coefficient that converts off-balance sheet exposures into credit exposure equivalents. It reflects the likelihood of an off-balance sheet item becoming an on-balance sheet item.
- **Slotting Criteria**: A set of predefined categories used for specialised lending exposures. These categories classify exposures based on qualitative assessments of the borrower's risk profile.
- **Validation Framework**: A comprehensive set of processes and procedures established by an institution to ensure the integrity and appropriateness of its risk models, including regular testing and review.

- **Internal Ratings-Based (IRB) Approach**: A method under the Basel Accords allowing banks to use their own estimated risk parameters for calculating regulatory capital requirements for credit risk.
- **Roll-out and Permanent Partial Use (PPU)**: The phased implementation of IRB models across different portfolios (roll-out) and the ongoing use of standardised approaches for certain exposures where IRB models are not applied (PPU).
- **Model Change Management**: The processes and controls involved in making changes to risk models, including documentation, approval, and implementation procedures to ensure model integrity over time.
- **Internal Validation Function**: An independent unit within a financial institution responsible for validating internal models. It assesses model performance, monitors limitations, and ensures compliance with regulatory requirements.
- **Statistical Tests**: Procedures for testing hypotheses about model parameters or distributions, considering limitations such as sample size and underlying assumptions.
- **Supervisory Review and Evaluation Process (SREP)**: A process whereby regulatory authorities, such as the ECB, assess an institution's strategies, processes, and risks to determine if it holds adequate capital and implements proper risk management.
- **Obligations and Recommendations**: Terms used in supervisory feedback. *Obligations* are mandatory actions that an institution must take to comply with regulatory requirements, while *Recommendations* are suggested improvements that are not legally binding.
- **Defaulted Assets**: Exposures where the borrower is unlikely to pay their credit obligations in full without recourse to actions such as the realization of collateral, or the borrower is past due more than 90 days on any material credit obligation.
- **Risk Parameters**: Quantitative measures used in credit risk modelling, including PD, LGD, EAD, and CCF, which quantify different aspects of credit risk.
- **Validation Tools**: Specific tests and analyses employed to evaluate and monitor the performance and appropriateness of risk models, such as back-testing, benchmarking, and sensitivity analyses.
- **Areas of Investigation**: The specific components or characteristics of a model that are examined during the validation process, including data quality, model assumptions, and computational processes.
- **Model Integrity**: The degree to which a model is sound, reliable, and operates as intended without errors or inconsistencies.
- **Modelling Assumptions**: The underlying premises accepted as true without proof in the development of a model, which can significantly impact model outcomes and must be validated for appropriateness.

- **Limitations of Statistical Tests**: Recognizing that statistical tests have constraints, such as sensitivity to sample size, data quality, and adherence to underlying assumptions, which can affect the interpretation of results.
- **Use Test**: A regulatory requirement ensuring that internal risk estimates play an essential role in the institution's risk management and decision-making processes, not just in regulatory capital calculations.
- **Qualitative Assessments**: Evaluations based on non-quantitative factors, such as expert judgment or qualitative criteria, often used in conjunction with quantitative measures in model validation.
- **Back-testing**: A validation technique where model predictions are compared against actual outcomes to assess predictive accuracy.
- **Benchmarking**: Comparing model outputs against external or internal reference models to identify discrepancies or areas for improvement.
- **Sensitivity Analysis**: Examining how changes in model inputs or assumptions affect outputs, used to identify which variables have the most significant impact on model results.
- **Regulatory Technical Standards (RTS)**: Detailed requirements developed by European Supervisory Authorities to enhance the harmonization and enforcement of regulations within the European Union.
- **European Central Bank (ECB) Instructions**: Guidelines and requirements issued by the ECB for institutions under its supervision, outlining expectations for compliance, model validation, and risk management practices.

2 Model Risk Management Principles

Model risk management is a critical aspect of financial institutions' overall risk management framework. It ensures that models are developed, implemented, and used appropriately, thereby mitigating potential adverse effects on the institution's financial health and compliance status.

- **Governance:** Establishing robust governance structures is essential. Clear definition of roles and responsibilities within the model risk management framework ensures accountability and effective oversight. For instance, specifying which units are responsible for independent reviews of risk estimates enhances transparency and objectivity in model assessments.
- **Model Risk Management Policy:** A comprehensive written policy is fundamental. This policy should define what constitutes a model, articulate the institution's interpretation of model risk, and describe the model risk framework, including its components such as model governance, risk control functions, validation functions, and internal audit. This documentation ensures consistent understanding and application of model risk principles across the institution.
- **Assessment and Measurement of Model Risk:** Implementing guidelines and methodologies for both qualitative and quantitative assessment of model risk is crucial. Institutions should regularly evaluate models to identify and measure areas of uncertainty and deficiencies, considering their materiality. This includes assessing data deficiencies, potential model misuse, and implementation issues that could impact model performance.
- **Data Quality:** High-quality data underpins reliable model outputs. Institutions must ensure the accuracy, completeness, and relevance of data used in model development and validation. Addressing data deficiencies proactively reduces model risk and enhances the credibility of model results.
- **Documentation:** Thorough documentation of all aspects of model risk management is imperative. This includes model development processes, validation activities, and policies and procedures. Proper documentation facilitates transparency, aids in internal and external reviews, and supports regulatory compliance.
- **Communication and Reporting:** Effective procedures for both internal and external communication are necessary. Regular reporting on model performance, risks, and mitigation strategies keeps stakeholders informed and engaged. Clear communication channels foster a culture of risk awareness and prompt action when issues arise.
- **Regulatory Standards Compliance:** Adherence to regulatory standards is non-negotiable. Institutions must stay informed about relevant regulations and ensure that their model risk management practices meet or exceed these requirements. Compliance not only satisfies legal obligations but also promotes best practices within the institution.

By embracing these principles, financial institutions can enhance their model risk management frameworks, leading to more reliable decision-making processes and strengthened overall risk management. This proactive approach not only safeguards the institution's interests but also contributes to the stability and integrity of the financial system as a whole.

2.1 Risk Management Frameworks

Effective model risk management is vital for financial institutions to reduce potential losses and prevent the underestimation of own funds requirements resulting from flaws in the development, implementation, or use of models. To mitigate these risks, institutions should establish a comprehensive model risk management framework that enables them to identify, understand, and manage model risk across all internal models within the group. This framework aligns with Basel and other regulatory requirements and should comprise at least the following components:

- **Model Register:** A detailed register of the institution's internal models, facilitating a holistic understanding of their application and use. This register provides the management body and senior management with a comprehensive overview of all models in place, as described in paragraph 6.
- **Defined Roles and Responsibilities:** Clear definitions of roles and responsibilities within the model risk management framework. This includes specifying which units are in charge of or involved in independent reviews of risk estimates, ensuring accountability and effective oversight.
- **Guidelines and Methodologies:** Established guidelines and methodologies for the qualitative and quantitative assessment and measurement of the institution's model risk. These guidelines ensure consistent evaluation and enable effective management of model performance and associated risks.

By implementing these components, institutions can effectively align their model risk management practices with regulatory expectations. A robust framework not only enhances the institution's ability to manage model risk but also supports compliance with Basel regulations and other supervisory requirements, thereby contributing to the overall stability and integrity of the financial system.

2.2 Model Governance and Oversight

Effective model governance and oversight are critical components of a financial institution's model risk management framework. Clear delineation of roles and responsibilities ensures that models are developed, validated, approved, and used appropriately, mitigating potential risks arising from model inaccuracies or misuse.

Institutions should establish a comprehensive framework that defines the roles and responsibilities of all parties involved in the model lifecycle. The key stakeholders and their responsibilities include:

- **Model Developers**: Responsible for the design, development, and initial testing of models. They ensure models are built according to sound methodological principles and in alignment with the institution's objectives. Developers should thoroughly document the technical aspects of the model, including methodologies and assumptions.
- **Model Validators**: Conduct independent reviews of models to assess their conceptual soundness, computational accuracy, and performance. Validators should be independent of the development process to maintain objectivity. Their duties include verifying model assumptions, testing implementation, and reviewing documentation.
- **Model Users**: Utilize models for decision-making processes. Users must understand the model's purpose, assumptions, limitations, and proper usage guidelines. They should follow documented instructions and report any issues or anomalies encountered during use.
- **Internal Audit**: Provides independent assurance on the effectiveness of the model risk management framework. Internal audit evaluates compliance with policies and procedures, reviews the governance processes, and ensures that model risk is adequately managed across the institution.
- **Senior Management**: Holds ultimate responsibility for model risk management. Senior management should define clear roles, authorities, and responsibilities within the model governance framework. They are responsible for approving models, allocating resources, and fostering a culture that emphasizes the importance of effective model risk management.

An essential component of model governance is maintaining a comprehensive **model inventory**. This inventory should include all models used within the institution, detailing:

- Model purpose and scope
- Ownership and accountability
- Development and validation history
- Approval status and dates
- Performance metrics and review schedules

The model inventory serves as a central repository, facilitating oversight and ensuring transparency throughout the model lifecycle.

Model approval processes must be clearly defined and enforced. Before deployment, models should undergo rigorous validation and receive formal approval from designated authorities. The approval process typically involves:

1. **Submission of Model Documentation**: Developers provide comprehensive documentation covering technical specifications, methodologies, assumptions, and user instructions.

2. **Independent Validation**: Validators perform an in-depth review, including testing model implementation and assessing performance against established criteria.

3. **Management Review and Approval**: Senior management or a designated committee reviews validation findings and approves the model for use, ensuring alignment with the institution's risk appetite and strategic objectives.

In the event of identified issues or when certain thresholds are breached, institutions should have established **escalation procedures**. These procedures ensure timely communication of model-related risks to appropriate levels of management. Key aspects include:

- **Defined Escalation Paths**: Clearly outlined channels for reporting model issues, including who should be notified at each escalation level.

- **Triggering Events**: Specific criteria that necessitate escalation, such as significant model performance deterioration or discovery of critical errors.

- **Resolution Protocols**: Steps to address and remediate issues, assignment of responsibilities, and timelines for action.

Institutions should also prioritize **training and awareness programs** tailored to the roles of different stakeholders. These programs may include workshops, seminars, or dedicated training sessions on model governance and risk management practices. Such initiatives ensure that all parties, from model developers to senior management, possess the necessary knowledge and understanding to fulfill their responsibilities effectively.

By defining clear roles and responsibilities and establishing robust processes for model inventory management, approval, and escalation, institutions reinforce their commitment to sound model governance. This comprehensive approach enhances transparency, accountability, and the overall integrity of the model risk management framework.

2.3 Data Quality Impact

In the realm of model validation within finance, the quality of data is paramount. High-quality data ensures that models accurately reflect real-world scenarios, leading to reliable predictions and sound decision-making. Conversely, poor data quality can compromise model integrity, resulting in flawed outcomes and potential regulatory non-compliance.

To emphasize the importance of data quality, it is essential to consider all relevant data quality dimensions:

- **Completeness**: Ensuring all necessary data is present, leaving no gaps that could skew model results.

- **Accuracy**: Verifying that data correctly represents the real values, minimizing errors and inaccuracies.
- **Consistency**: Maintaining uniformity across datasets, so that data does not conflict within different parts of the model.
- **Timeliness**: Utilizing up-to-date data to reflect the most current conditions affecting the model.
- **Uniqueness**: Avoiding duplicate entries, which can distort analyses and outcomes.
- **Validity**: Confirming that data conforms to the required formats and standards, founded on an adequate system of classification rigorous enough to compel acceptance.
- **Availability**: Ensuring data is readily accessible when needed for analysis and reporting.
- **Traceability**: Keeping a clear record of data sources and transformations for auditing and verification purposes.

An effective data quality framework should cover all these dimensions. Such a framework should clearly define policies, roles, and responsibilities in data processing and data quality management. This includes establishing data quality standards and implementing processes to monitor and maintain data integrity.

Investigations into data management practices have revealed significant shortcomings:

- **Control Framework Issues**: 77% of investigations found insufficiencies or lack of controls on relevant data elements of the models under investigation. Institutions struggled to prove the quality of data used, highlighting weaknesses in their control frameworks.
- **Data Quality Systems and Processes**: 56% of investigations reported shortcomings in data quality systems, procedures, and processes. Issues included inadequate coverage of models within the scope of investigations and problems related to the Internal Ratings-Based (IRB) data supporting the models or key elements of the IRB data cycle.

These findings underscore the critical impact of data quality on model validation. Without robust data quality systems:

- Models may yield inaccurate or misleading results due to incomplete or erroneous data.
- Regulatory compliance may be compromised, leading to potential penalties and reputational damage.
- Decision-making processes may be flawed, affecting strategic and operational outcomes.

Treatment of Missing Data

Handling missing data is a crucial aspect of data quality management. Missing data can occur due to various reasons, such as system errors, manual input mistakes, or inaccessible data sources. The treatment of missing data involves:

- **Identification**: Detecting missing data points within the dataset.
- **Analysis**: Assessing the extent and patterns of missingness to understand potential impacts on the model.
- **Imputation**: Employing statistical methods to estimate and replace missing values where appropriate.
- **Model Adjustment**: Modifying models to account for missing data, such as using models robust to missingness or incorporating missingness indicators.

Effective management of missing data enhances the completeness dimension of data quality, ensuring that models are built on a solid foundation.

Conclusion

Data quality directly impacts the efficacy of model validation in finance. By comprehensively addressing all data quality dimensions—completeness, accuracy, consistency, timeliness, uniqueness, validity, availability, and traceability—institutions can strengthen their models' reliability and integrity. Implementing a robust data quality framework with clearly defined policies and responsibilities is essential. This commitment to data quality not only enhances model performance but also supports regulatory compliance and fosters trust among stakeholders.

2.4 Model Documentation

Thorough model documentation is a critical component in the lifecycle of financial models. Proper documentation ensures that a qualified third party can independently understand the methodology, assumptions, limitations, and usage of a model, and can replicate its development and implementation. This level of transparency is essential not only for internal governance but also for satisfying regulatory requirements and facilitating supervisory reviews.

Documentation should be kept up to date, and institutions should retain documents for an appropriate period, taking into account legal or regulatory retention periods. An effective documentation framework supports consistency, accountability, and the ability to track changes over time.

Institutions should define clear principles and guidelines for model documentation, encompassing governance of the documentation process itself. The scope of the documentation should be tailored to the type of model but should include at least the following elements:

- **Technical Aspects:** Detailed descriptions of the model's methodology and underlying assumptions, including any theoretical foundations.

- **Data Processes:** Information on data preparation, sources, quality checks, and handling procedures used during development and calibration.
- **User Instructions:** Guidance for model users on operation, implementation in internal processes, and interpretation of results.
- **Performance and Validation:** Results from validation activities, including implementation testing, performance metrics, and any limitations identified.

Adequate controls over the institution's register of internal models, along with an inventory of associated documentation, should be in place. This involves conducting regular reviews—at least annually—to ensure completeness and consistency. Policies for document management should clearly state roles and responsibilities for approving documents, applying changes, and communicating updates internally. Additionally, institutions should establish policies regarding:

- **Archiving and Maintenance:** Procedures for proper archiving of documentation and maintenance of historical records.
- **Access Permissions:** Controls over who can access documentation to protect sensitive information.
- **Completeness and Consistency Assessments:** Regular evaluations to ensure all necessary information is documented and consistently presented.

The model life cycle generally includes the following steps:

1. **Development:** Crafting the initial model framework, including data preparation.
2. **Calibration:** Adjusting model parameters to fit historical data, with further data preparation as needed.
3. **Validation:** Independent assessment of the model's performance, assumptions, and limitations.
4. **Supervisory Approval (if necessary):** Obtaining approval from relevant regulatory bodies.
5. **Implementation in Internal Processes:** Integrating the model into business operations and decision-making processes.
6. **Application and Review of Estimates:** Ongoing use of the model and periodic review of its outputs and effectiveness.

Each phase should be thoroughly documented to provide a comprehensive audit trail and facilitate effective risk management.

Key Elements of Model Documentation

- **Model Development Documents:** These should capture the rationale behind the model's design, including:

  - Objectives and intended use of the model.
  - Detailed methodology and theoretical justification.
  - Assumptions and limitations inherent in the model.
  - Data sources, selection criteria, and processing steps.
  - Results from developmental testing and initial performance assessments.

- **Validation Reports:** Documentation of independent validation activities, encompassing:
  - Summary of validation scope and methodologies employed.
  - Findings from performance tests, including back-testing and benchmarking.
  - Identification of model risks and limitations.
  - Recommendations for remediation or enhancements.

- **Ongoing Monitoring Reports:** Periodic assessments that track the model's performance over time, addressing:
  - Updates on model outputs and key performance indicators.
  - Analysis of any deviations or unexpected results.
  - Record of incidents and exception handling.
  - Documentation of changes in external conditions impacting the model.

By adhering to robust documentation practices, institutions not only comply with regulatory expectations but also enhance the integrity and reliability of their modeling activities. Comprehensive documentation supports effective communication among stakeholders, facilitates training for new users, and underpins the institution's capability to respond to audits or regulatory inquiries.

2.5 Regulatory Standards

Credit risk model validation plays a pivotal role in ensuring the robustness and reliability of financial institutions' internal models. To maintain financial stability and protect against systemic risks, regulatory bodies have established comprehensive standards and guidelines governing the development, implementation, and validation of these models. This section outlines the key regulatory frameworks relevant to credit risk model validation, including the Basel Accords, SR 11-7, Current Expected Credit Loss (CECL), International Financial Reporting Standard 9 (IFRS 9), and specific requirements set forth by the European Central Bank (ECB).

Basel Accords (Basel II and Basel III): Developed by the Basel Committee on Banking Supervision, the Basel Accords provide a set of international banking regulations aimed at enhancing the quality of capital and promoting robust risk management practices. The Capital Requirements Regulation (CRR), which transposes Basel standards into European Union law, mandates that credit institutions must ensure their internal models for credit risk are subject to a rigorous validation process. Specifically, Articles 174(d), 185, and 188 of the CRR stipulate the requirements for model validation, emphasizing the need for institutions to:

- Verify the *adequacy*, *robustness*, and *reliability* of internal estimates used to calculate own funds requirements.
- Develop validation processes that comply with regulatory standards, allowing for effective comparisons across models and institutions.
- Ensure that validation reports enable senior management to understand model performance and identify potential weaknesses.

SR 11-7: Issued by the Board of Governors of the Federal Reserve System in 2011, SR 11-7 provides comprehensive guidance on model risk management for banking organizations. It outlines the essential components of an effective model risk management framework, including:

- *Model development, implementation, and use*: Ensuring models are conceptually sound and appropriately implemented with well-defined applications.
- *Model validation*: Conducting independent and rigorous validation processes to assess model performance and limitations.
- *Governance, policies, and controls*: Establishing strong governance structures with clear responsibilities and oversight mechanisms.

SR 11-7 emphasizes that validation should be an ongoing process, incorporating both quantitative and qualitative assessments to manage model risk effectively.

Current Expected Credit Loss (CECL): Introduced by the Financial Accounting Standards Board (FASB) in the United States, CECL represents a fundamental shift in how financial institutions recognize credit losses. Under CECL, institutions are required to estimate *expected credit losses* over the life of financial assets, rather than recognizing losses when they become probable. This forward-looking approach necessitates the development and validation of robust credit risk models capable of:

- Incorporating reasonable and supportable forecasts.
- Handling a range of macroeconomic scenarios.
- Adjusting for changes in asset quality and portfolio composition.

International Financial Reporting Standard 9 (IFRS 9): IFRS 9, issued by the International Accounting Standards Board (IASB), sets forth the requirements for recognition and measurement of financial instruments, including the impairment of financial assets. Similar to CECL, IFRS 9 introduces an *expected credit loss* model, requiring institutions to account for credit losses from the time a financial instrument is first recognized. Key aspects of IFRS 9 related to model validation include:

- Developing models that can estimate expected credit losses over varying time horizons.

- Incorporating forward-looking information and multiple economic scenarios.
- Regularly validating models to ensure accuracy and compliance with accounting standards.

European Central Bank (ECB) Requirements: The ECB mandates specific guidelines for credit institutions within the Eurozone, aiming to harmonize supervisory practices and ensure financial stability. According to Regulation (EU) No 575/2013 (CRR), institutions must develop validation processes that:

- Verify the *overall adequacy*, *robustness*, and *reliability* of internal models used for calculating own funds requirements.
- Include various types of analyses on key modelling assumptions as part of a regular validation schedule, as considered best practice by the ECB.
- Cover necessary testing to ensure model integrity and appropriateness of assumptions, in accordance with Article 292(6)(a) of the CRR.

The ECB emphasizes the importance of:

- Producing validation reports that adhere to regulatory standards and allow for effective oversight by senior management.
- Ensuring validation processes comply with the relevant Regulatory Technical Standards (RTS) and facilitate comparisons across models and institutions.
- Focusing on key modelling assumptions, including data quality, estimation techniques, and calibration methods.

Summary: Adherence to these regulatory standards is essential for financial institutions to maintain sound credit risk management practices. A robust model validation framework not only ensures compliance but also enhances the institution's ability to manage risk effectively in an evolving financial landscape. By aligning with international frameworks such as the Basel Accords, adhering to national guidelines like SR 11-7, and fulfilling accounting standards under CECL and IFRS 9, institutions can uphold the integrity of their credit risk models. Furthermore, meeting the ECB's specific requirements ensures harmonization within the Eurozone, contributing to the overall stability of the financial system.

2.6 Ethical Considerations

The implementation of the Internal Ratings-Based (IRB) Approach in financial institutions raises several ethical considerations that are critical to maintaining the integrity of financial systems. Key ethical aspects include transparency, prevention of biased practices, responsible outsourcing, data integrity, and adherence to regulatory guidelines.

2.6.1 Transparency and Disclosure

Transparency is paramount in fostering trust and accountability within the financial industry. Institutions should support guidelines, such as those proposed by the European Banking Authority (EBA) and the Basel Committee, that enhance disclosure requirements related to the IRB Approach. Prioritizing current disclosure requirements ensures that stakeholders are well-informed about the risk profiles and modeling practices of institutions. Establishing a clear timetable for implementing such guidelines demonstrates a commitment to ethical transparency and allows for harmonized adoption across the industry.

2.6.2 Preventing *Cherry Picking*

An ethical concern in the application of the IRB Approach is the potential for *cherry picking*, where institutions might selectively use internal models for exposures that result in lower capital requirements while applying standardized approaches elsewhere. To ensure the appropriate use of the IRB Approach without excessive *cherry picking* concerns, institutions should:

- Adopt a harmonized application of internal models across all relevant exposure classes.
- Ensure consistent treatment of similar exposures to prevent biases.
- Align modeling practices with the institution's business model and operational capacity.

By addressing these factors, institutions uphold ethical standards and contribute to the stability of the financial system.

2.6.3 Responsible Outsourcing Practices

Outsourcing IRB-related tasks introduces ethical considerations regarding accountability and control over risk assessment processes. All outsourcing arrangements should be governed by formal and comprehensive contracts or similar documented agreements, adhering to the proportionality principle. In cases of internal outsourcing within the same group, service level agreements (SLAs) or other written agreements may be sufficient, depending on the criticality of the tasks outsourced.

Institutions should consider:

- The importance of the exposure classes or types of exposures to the institution's business model.
- Data availability and quality for accurate risk assessment.
- Operational capacity and staffing to manage outsourced tasks effectively.

- Length of experience with the outsourced activities.
- Homogeneity in the treatment of similar exposures to maintain consistency.

By carefully evaluating these aspects, institutions can ensure that outsourcing does not compromise ethical standards or risk management effectiveness.

2.6.4 Data Integrity and Representativeness

Maintaining the integrity and representativeness of data used in IRB models is an ethical imperative, especially in the context of extraordinary events such as the COVID-19 pandemic. Institutions should adhere to principles that ensure:

- Data reflects current and relevant risk factors.
- Models are adjusted to account for atypical data periods.
- Ongoing validation and review of models in light of new information.

Reference materials, such as the document "Principles on Representativeness of COVID-19 Impacted IRB Relevant Data," provide guidance on maintaining data integrity during unprecedented times.

2.6.5 Regulatory Compliance and Ethical Responsibility

Adhering to regulatory guidelines is both a legal obligation and an ethical responsibility. Institutions should proactively engage with regulatory bodies to support the development and implementation of guidelines that enhance the IRB Approach. This includes:

- Supporting EBA Guidelines on disclosure requirements.
- Aligning with proposals from the Basel Committee.
- Prioritizing guidelines that have the most significant impact on transparency and risk management.

Timely adoption of these guidelines demonstrates an institution's commitment to ethical practices and contributes to the overall health of the financial system.

3 PD Discriminatory Power Tests

Analyses of discriminatory power for Probability of Default (PD) models are essential to ensure that the ranking of obligors or facilities resulting from the rating methodology appropriately separates riskier entities from less risky ones. A well-calibrated PD model should effectively differentiate between obligors with high default risk and those with low default risk, thereby enhancing risk assessment and decision-making processes.

Similarly, analyses of discriminatory power for Loss Given Default (LGD) and Credit Conversion Factor (CCF) models are designed to ensure that these models can discriminate between facilities with high and low values of LGD or CCF. Effective discrimination in these models contributes to more accurate loss estimations and better capital allocation.

For certain tests assessing discriminatory power, it is necessary to estimate the variance of the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve or the covariance of the AUCs from two different rating assignments for a given portfolio during the observation period. These statistical measures provide insights into the consistency and reliability of the model's discriminatory capabilities over time.

The validation tools discussed in this section focus on comparing the discriminatory power of a PD model in its current state with the discriminatory power it exhibited during the initial validation at the time of development. By examining changes in the AUC over time, practitioners can assess whether the model's ability to discriminate between different levels of default risk has improved, remained stable, or deteriorated.

An essential criterion for PD model validation is that the assigned ratings correspond appropriately to the estimated PDs. Specifically, if obligor a is assigned a better (higher) rating than obligor b, it should imply that the estimated PD for a is lower than that for b. This relationship confirms that the model assigns lower risk ratings to obligors with lower probabilities of default, reflecting good discriminatory power.

3.1 Introduction to Discriminatory Power

Discriminatory power is a fundamental concept in credit risk modeling that assesses the ability of a rating system or predictive model to correctly distinguish between riskier and less risky customers or facilities. It ensures that the ranking derived from the rating process appropriately separates entities based on their likelihood of default or loss severity.

In the context of Probability of Default (PD) models, the analysis of discriminatory power aims to verify that the methodology effectively differentiates between obligors who are more likely to default and those who are less likely. This is crucial for accurate risk assessment and decision-making in credit approvals, pricing, and portfolio management.

For Loss Given Default (LGD) and Credit Conversion Factor (CCF) models, discriminatory power evaluates the models' ability to distinguish between facilities with high and low LGD or CCF values. This differentiation is essential for estimating potential losses and exposures accurately.

Assessing discriminatory power is important because:

- It validates the effectiveness of risk models in ranking customers or facilities according to risk.
- It supports regulatory compliance by ensuring models meet required performance standards.
- It enhances risk management practices by improving the accuracy of credit risk evaluations.

Various measures exist to quantify discriminatory power, with the Area Under the Curve (AUC) being one of the most widely used. The AUC provides a single metric that summarizes the model's ability to rank-order risk correctly across all possible classification thresholds. For PD models, the AUC can be defined and calculated using the Mann-Whitney U statistic. In the case of LGD and CCF models, a generalized version of the AUC is employed to handle multi-class problems effectively.

Regular analysis of discriminatory power is essential for model validation and monitoring. Comparing the current discriminatory power of a PD model to its performance during initial development helps identify any degradation over time. This comparison ensures that the model remains robust and continues to provide reliable risk assessments.

In summary, discriminatory power is a critical attribute of credit risk models that affects the accuracy of risk ranking and the effectiveness of risk management strategies. By ensuring that models have strong discriminatory power, financial institutions can make more informed decisions and maintain compliance with regulatory standards.

3.2 Key Tests

The analysis of discriminatory power is essential to validate that the ranking of customers resulting from the rating process effectively distinguishes between riskier and less risky customers. This assessment ensures that credit risk models, such as Probability of Default (PD), Loss Given Default (LGD), and Credit Conversion Factor (CCF) models, are robust in differentiating levels of risk.

Several measures exist to evaluate discriminatory power, with the Area Under the Curve (AUC) being a widely accepted metric. The AUC, defined in terms of the Mann-Whitney U statistic, provides a quantifiable means to assess how well the model separates different risk categories without relying on specific distributional assumptions.

For PD models, the key tests focus on verifying that the ranking of obligors or facilities appropriately reflects their likelihood of default. This involves:

- **Calculating the AUC**: A high AUC value indicates strong discriminatory power in separating defaulted from non-defaulted obligors.
- **Assessing the Mann-Whitney U statistic**: This non-parametric test evaluates whether the scores assigned to defaulted obligors are statistically higher than those for non-defaulted obligors.

Similarly, for LGD models, the analysis aims to ensure the model effectively discriminates between facilities with high and low loss rates. The key tests include:

- **Using the generalised AUC**: This measure extends the traditional AUC to handle multi-class problems inherent in LGD modeling, where losses can take on a range of values.
- **Evaluating segmentation**: Testing the model across different segments to verify consistent discriminatory power throughout various facility types.

For CCF models, which estimate exposure at default, the key tests are designed to confirm the model's ability to distinguish between facilities with high and low CCF values. This involves:

- **Analyzing the ranking quality**: Ensuring that the model assigns higher CCF estimates to facilities that are more likely to draw down their committed but undrawn lines.
- **Comparing predicted and actual CCF values**: Assessing whether the model's predictions align with observed outcomes across different facility groups.

Implementing these key tests allows institutions to validate the discriminatory power of their risk models effectively. By focusing on the AUC and its generalisations, as well as utilizing statistical tests like the Mann-Whitney U statistic, model validators can ensure that the models provide meaningful risk differentiation in compliance with regulatory standards.

ROC Curve and AUC The Receiver Operating Characteristic (ROC) curve is a fundamental tool for evaluating the discriminatory power of binary classification models in finance. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, providing a comprehensive view of the model's ability to distinguish between classes.

To calculate the ROC curve, one must:

1. Rank the predicted probabilities or scores from the model in descending order.
2. For each unique threshold, classify observations as positive if the predicted probability exceeds the threshold, and negative otherwise.
3. Compute the TPR and FPR for each threshold:
    - TPR: The proportion of actual positives correctly identified (also known as sensitivity).
    - FPR: The proportion of actual negatives incorrectly classified as positive.
4. Plot the TPR against the FPR to visualize the ROC curve.

The Area Under the ROC Curve (AUC) is a single scalar value summarizing the performance of the model across all thresholds. An AUC of 1.0 indicates perfect discrimination, while an AUC of 0.5 suggests no discrimination (equivalent to random guessing). The

AUC is calculated using methods such as the trapezoidal rule to approximate the area under the ROC curve.

Interpretation:

The ROC curve and AUC provide insights into the model's ability to rank-order risk correctly. A higher AUC indicates better model performance in terms of discriminatory power. In the context of regulatory compliance, it is crucial to assess not only the model's performance on the development dataset but also on Out-of-Sample (OOS) and Out-of-Time (OOT) datasets to ensure robustness and avoid overfitting.

Visualization:

Visualizing the ROC curve allows practitioners to evaluate the trade-off between sensitivity and specificity at different thresholds. It is considered best practice to plot ROC curves for both the development and validation datasets, including OOS and OOT samples. This comparison helps in identifying any degradation in model performance over time or across different samples.

An example of how to plot the ROC curve using Python is provided below:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc

# True binary labels
y_true = np.array([...])  # Replace with actual true labels
# Predicted probabilities or scores from the model
y_scores = np.array([...])  # Replace with predicted probabilities

# Compute False Positive Rate (fpr), True Positive Rate (tpr), and
    thresholds
fpr, tpr, thresholds = roc_curve(y_true, y_scores)
# Compute Area Under the Curve (AUC)
roc_auc = auc(fpr, tpr)

# Plot the ROC curve
plt.figure()
plt.plot(fpr, tpr, label='ROC curve (AUC = {:.2f})'.format(roc_auc))
# Plot the diagonal reference line
plt.plot([0, 1], [0, 1], 'k--')
# Set the limits for x and y axes
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
# Label the axes
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# Add a title
plt.title('Receiver Operating Characteristic')
# Add a legend
plt.legend(loc='lower right')
# Display the plot
plt.show()
```

Creating Confidence Intervals:

To assess the statistical significance of the AUC, confidence intervals can be constructed using methods such as bootstrapping or the DeLong method. Confidence intervals provide a range within which the true AUC is expected to lie with a certain level of confidence (typically 95%). This is particularly important in cases of data scarcity, where small sample sizes can lead to greater uncertainty in the estimates.

Regulatory Considerations:

According to regulatory guidelines, institutions should include references to measures of discriminatory power like the AUC in their internal validations, regardless of whether the same techniques are used in the institution's model development. It is imperative that the validation function has access to empirical assessments of the model's performance on OOS and OOT samples conducted by the Credit Risk Control Unit (CRCU). This ensures a complete and independent assessment of the model.

However, challenges may arise in performing validations based on OOS and OOT samples in the context of data scarcity. Institutions must navigate these challenges carefully, possibly by complementing analyses with additional validation tests based on the full dataset, while maintaining the independence of the validation process.Gini Coefficient

The Gini coefficient is a measure used to assess the discriminatory power of the estimated Loss Given Default (LGD_E) models by evaluating how well they predict the ordering of realized LGD values (LGD_R). In this context, the Gini coefficient is calculated based on a 12×12 contingency table, which represents all possible combinations of discretized LGD_E segments (rows) and LGD_R segments (columns).

Let a_{ij} denote the observed frequency in cell (i, j) of the contingency table, where:

- i represents the segment for LGD_E,
- j represents the segment for LGD_R.

Define the following totals:

$$r_i = \sum_j a_{ij} \quad \text{(total frequency for row } i\text{)},$$
$$c_j = \sum_i a_{ij} \quad \text{(total frequency for column } j\text{)},$$
$$F = \sum_i \sum_j a_{ij} \quad \text{(total frequency of all observations)}.$$

To compute the Gini coefficient, we calculate the values of P and Q as follows:

- P (the number of agreements) is twice the sum of the frequencies where both the estimated and realized LGD indices are either greater or smaller than a given combination. Mathematically, it represents the total frequency of observations where the ordering of LGD_E and LGD_R agree.
- Q (the number of disagreements) is twice the sum of the frequencies where the estimated and realized LGD indices do not agree in ordering. Specifically, it counts the observations where one index is greater and the other is smaller than the given combination.

Using these definitions, the Gini coefficient (G) is calculated using the formula:

$$G = \frac{P - Q}{P + Q}. \tag{1}$$

Alternatively, the Gini coefficient is related to Somers' D statistic, which in this context is estimated as:

$$\text{Somers' } D(C|R) = \frac{P - Q}{F^2 - \sum_i r_i^2}, \tag{2}$$

where $C|R$ indicates that the column variable (LGD_R) is regarded as dependent on the row variable (LGD_E). The Gini coefficient is then obtained by:

$$G = 2 \times \text{Somers' } D(C|R). \tag{3}$$

Summary of Steps to Calculate the Gini Coefficient:

1. Construct a 12×12 contingency table of discretized LGD_E and LGD_R segments.
2. Calculate the observed frequencies a_{ij} for each cell (i, j).
3. Compute the row totals r_i, column totals c_j, and total frequency F.
4. Determine P and Q based on the number of agreements and disagreements in ordering between LGD_E and LGD_R.
5. Apply the formula to calculate the Gini coefficient G.

Reference: Göktaş, A. and İşçi, Ö. (2011), "A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation", *Metodološki zvezki*, Vol. 8(1), pp. 17-37.

Kolmogorov-Smirnov (KS) Statistic The Kolmogorov-Smirnov (KS) Statistic is a non-parametric test used to compare an empirical distribution with a reference probability distribution or to compare two empirical distributions. In model validation within finance, it serves as a tool to assess the goodness-of-fit between observed data and a modeled distribution.

To calculate the KS Statistic, follow these steps:

1. **Sort the Data:** Arrange the sample data in ascending order to obtain a sequence $x_1, x_2, \ldots, x_n$.
2. **Compute Empirical Cumulative Distribution Function (ECDF):** For each data point x_i, calculate the ECDF:

   $$F_n(x_i) = \frac{\text{Number of observations } \leq x_i}{n} = \frac{i}{n}$$

   where n is the total number of observations.

3. **Determine the Theoretical Cumulative Distribution Function (CDF):** Calculate the theoretical CDF $F(x_i)$ at each x_i based on the reference distribution.

4. **Calculate the Differences:** Find the absolute differences between the ECDF and the theoretical CDF at each point:

$$D_i = |F_n(x_i) - F(x_i)|$$

5. **Find the KS Statistic:** The KS Statistic D_n is the maximum of these differences:

$$D_n = \max_{1 \leq i \leq n} D_i$$

6. **Assess the Significance:** Compare D_n against critical values from the KS distribution or compute the p-value to assess the significance of the observed difference.

A smaller KS Statistic indicates a closer agreement between the empirical data and the reference distribution. In regulatory compliance, this test helps validate whether the modeled distributions accurately represent the underlying risk factors.

Example in Python:

```
import numpy as np
from scipy import stats

# Sample data
data = np.array([...])  # Replace with actual data

# Sort the data
data_sorted = np.sort(data)
n = len(data_sorted)

# Compute ECDF
ecdf = np.arange(1, n+1) / n

# Define the reference CDF (e.g., standard normal distribution)
ref_cdf = stats.norm.cdf(data_sorted, loc=..., scale=...)  # Replace
    loc and scale with model parameters

# Calculate the KS Statistic
D_n = np.max(np.abs(ecdf - ref_cdf))

# Compute p-value
p_value = stats.kstest(data, 'norm', args=(..., ...)).pvalue  # Replace
     args with model parameters

# Output the results
print(f"KS Statistic: {D_n}")
print(f"P-value: {p_value}")
```

In this code:

- We import the necessary libraries. - The data is sorted to compute the ECDF. - The reference CDF is calculated using the presumed distribution (e.g., normal distribution with specified parameters). - The KS Statistic is the maximum absolute difference between the ECDF and the reference CDF. - The p-value is computed using SciPy's built-in KS test function.

Interpretation: A p-value below a chosen significance level (e.g., 0.05) indicates that there is a statistically significant difference between the empirical distribution and the reference distribution, suggesting that the model may not adequately capture the observed data's characteristics.

Cumulative Accuracy Profile (CAP) The Cumulative Accuracy Profile (CAP) is a graphical tool used to assess the predictive ability of credit risk models, specifically the effectiveness of the Credit Conversion Factor (CCF) in forecasting Exposure at Default (EAD). It evaluates how well the model discriminates between facilities with higher and lower risks of default, ensuring that the CCF risk parameter facilitates a good prediction of EAD.

To calculate the CAP, follow these steps:

1. **Construct the Sample:** Assemble a dataset of facilities whose recovery process has begun within the relevant one-year observation period. This includes collecting both the estimated CCFs to be back-tested and the actual outcomes for each facility. Facilities covered by an EAD approach may require a simplified analysis as per regulatory guidelines.

2. **Order the Facilities:** Rank the facilities in descending order based on their estimated CCFs or scores. Facilities with higher estimated CCFs are expected to contribute more to the EAD.

3. **Calculate Cumulative Proportions:**
   - Determine the cumulative number of facilities at each rank position.
   - Calculate the cumulative sum of actual EADs corresponding to these facilities.
   - Express these cumulative sums as percentages of the total number of facilities and the total actual EADs, respectively.

4. **Plot the CAP Curve:** On a graph:
   - The horizontal axis represents the cumulative percentage of facilities.
   - The vertical axis represents the cumulative percentage of actual EADs.
   - Plot points for each facility's cumulative percentages and connect them to form the CAP curve.

5. **Interpret the CAP Curve:**
   - An ideal model would show a CAP curve that rises sharply, indicating that a small percentage of facilities account for a large percentage of actual EADs.
   - A model with no predictive power would produce a CAP curve along the diagonal, signifying that the cumulative EAD increases proportionally with the number of facilities.

- The closer the CAP curve is to the ideal model's curve, the better the model's discriminatory power.

6. **Assess Predictive Ability:** Use the area under the CAP curve to quantify the model's performance. Comparing this area to that of the ideal and random models provides a measure of accuracy. Regulatory compliance may require calculating specific statistical measures, such as p-values using the cumulative distribution function of the Student's t-distribution, to assess the model's significance.

7. **Apply Regulatory Calibration:** According to Article 180(1)(a) of the CRR, perform explicit calibration at the grade or pool level based on individual estimates or scores. Assign facilities to intervals on a predefined master scale where the long-run average default rate matches the predefined default rate for that interval. Ensure that the calibration sample matches the sample used for assessing individual estimates.

By following these steps, the CAP provides a visual and statistical method to evaluate the predictive accuracy of CCF models. It ensures that the models are effectively distinguishing between different levels of credit risk and are aligned with regulatory expectations for robust risk parameter estimation.

Binned vs. Unbinned Tests In the realm of model validation and regulatory compliance in finance, statistical tests play a crucial role in assessing the performance and reliability of risk models. Two fundamental approaches to conducting these tests are **binned** and **unbinned** tests, each with its own advantages and considerations.

Binned Tests involve grouping continuous data into discrete intervals or *bins* before performing statistical analysis. This approach simplifies the data by aggregating individual observations within predefined ranges, making it easier to observe overall trends and patterns. Binned tests are particularly useful when dealing with large datasets or when the objective is to understand the distribution of variables across different segments. By focusing on the aggregated behavior within each bin, these tests can highlight deviations from expected patterns in specific intervals.

In contrast, **Unbinned Tests** analyze data at the individual observation level without any prior grouping. This method retains the full granularity of the data, allowing for a more detailed and precise statistical analysis. Unbinned tests are advantageous when the sample size is manageable or when the intricacies of individual data points are significant for the model's validation. They are more sensitive to subtle variations and can provide insights into the nuanced performance of the model across the entire dataset.

The choice between binned and unbinned tests depends on several factors:

- **Data Characteristics**: The nature of the dataset, including its size and the distribution of variables, influences the suitability of each testing approach.
- **Validation Objectives**: Whether the goal is to detect overall trends or to scrutinize individual observations affects the decision.

- **Regulatory Requirements**: Compliance standards may mandate specific testing methodologies or thresholds that favor one approach over the other.

For instance, in portfolio calibration (as highlighted in the background information), the number of time slices used can impact the comparability and representativeness of the calibration sample. Using only one point in time might necessitate binned tests to ensure the sample aligns with current portfolio characteristics. Conversely, incorporating all available time slices could benefit from unbinned tests to capture the full range of variability in obligor and transaction characteristics.

It's imperative for institutions to carefully select the appropriate testing method. As noted in regulatory observations, failures often occur when institutions do not ensure that all suitable validation analyses are performed with adequate quantitative thresholds. This includes essential tests such as back-testing, discriminatory power assessments, stability tests, and analyses of overrides.

In summary, both binned and unbinned tests have their place in model validation. Understanding their differences and applications enables institutions to effectively evaluate model performance, uphold regulatory compliance, and make informed decisions based on robust statistical evidence.

3.3 Benchmarking Discriminatory Power

Benchmarking the discriminatory power of credit risk models is crucial to ensure that the models effectively differentiate between varying levels of risk among obligors or facilities. This process involves comparing the model's ability to rank-order exposures from the riskiest to the least risky, both against internal standards and external benchmarks.

For Probability of Default (PD) models, analyses should confirm that the rating methodology appropriately separates riskier obligors from less risky ones. Similarly, for Loss Given Default (LGD) and Credit Conversion Factor (CCF) models, it is essential to demonstrate that the models can distinguish between facilities with high LGD or CCF values and those with low values.

Internal Benchmarking: The initial validation of discriminatory power should be performed using internal data. By evaluating the model's performance on data reflective of the institution's own portfolio, analysts can identify any weaknesses specific to their context. This internal assessment ensures that the model accurately captures the risk characteristics inherent in the institution's exposures.

External Benchmarking: Where available, external data sources provide an additional layer of validation. Utilizing external ratings as a challenger model is considered best practice when a sufficient number of such ratings are accessible. This comparison serves not as an absolute benchmark but as a tool to uncover potential weaknesses in the internal model's effectiveness. It helps in assessing whether the model adequately considers all relevant information compared to broader market standards.

In cases where data scarcity necessitates the use of external data in model development, performance assessments should include both internal and external data. Complementing internal evaluations with external data assessments enhances the robustness of the validation process, ensuring that the model maintains its discriminatory power across different datasets.

Overall, benchmarking against both internal and external references is a vital component of the model validation process. It ensures that credit risk models remain effective in ranking obligors or facilities according to their risk levels, thereby supporting more accurate risk measurement and management within financial institutions.

4 PD Calibration Tests

Probability of Default (PD) calibration tests are essential tools for assessing the accuracy and reliability of PD models used in credit risk management. Calibration ensures that PD estimates correspond to the observed long-run average default rates (DR), which is crucial for appropriate risk quantification and regulatory compliance.

According to regulatory guidelines (GLs), calibration should be conducted *before* the application of the PD floor or Margin of Conservatism (MoC). This requirement follows the same logic as for MoC; applying the PD floor prior to calibration can lead to less conservative PD estimates, potentially underestimating credit risk. Calibration prior to the PD floor ensures that the PD estimates are aligned with historical default experiences, providing a more accurate reflection of the underlying risk.

In practice, this requirement is largely adhered to, with approximately 75% of PD models (representing 79% of exposures) conducting calibration before applying the PD floor. However, there remains a significant portion—around 20% of PD models—that require a change in practice to comply with the GLs. Aligning these models with the guidelines enhances the consistency and conservatism of PD estimates across institutions.

PD calibration tests involve comparing the predicted PDs from the model to the actual observed default rates over a specified period. Several common tests and methods are used to assess calibration quality:

- **Binomial Test**: Evaluates whether the number of observed defaults is statistically consistent with the predicted PDs, assuming defaults follow a binomial distribution.
- **Hosmer-Lemeshow Test**: Divides the portfolio into groups based on predicted PDs and compares the observed and expected defaults in each group to assess the goodness-of-fit.
- **Calibration Plot**: A graphical tool that plots observed default rates against predicted PDs to visually assess calibration accuracy.
- **Brier Score**: A metric that quantifies the accuracy of probabilistic predictions by measuring the mean squared difference between predicted PDs and actual outcomes.

These tests help institutions identify discrepancies between predicted PDs and actual default rates, facilitating model adjustments to improve calibration. The ultimate aim is to ensure that, for each calibration segment, PD estimates correspond to the long-run average default rates relevant to the applied calibration method.

It's also important to acknowledge calibration types and methods beyond the standard approaches. Institutions may employ alternative calibration techniques tailored to specific portfolio characteristics or risk profiles. Incorporating diverse methods can enhance model performance but requires careful validation to ensure compliance with regulatory expectations.

In summary, PD calibration tests are a critical component of model validation in credit risk management. Conducting calibration before the application of the PD floor, as required by the GLs, promotes more conservative and accurate PD estimates. Regularly performing calibration tests and adjusting models accordingly ensures that PD estimates remain aligned with observed default experiences, thereby strengthening risk management practices and regulatory compliance.

4.1 Introduction to Calibration

In credit risk modeling, *calibration* refers to the process of adjusting model parameters to ensure that the model's outputs accurately reflect observed data and reliably predict future default events. The analysis of predictive ability, or calibration, aims to ensure that the Probability of Default (PD) parameter adequately predicts the occurrence of defaults—that is, that PD estimates constitute reliable forecasts of default rates.

Calibration is crucial for both regulatory reporting and internal risk management purposes. The calibration frequency is particularly relevant as part of the use test requirements set by Article 289 of the Capital Requirements Regulation (CRR). Regular calibration minimizes the risk of noncompliance with Article 292(2) and Article 289(5) of the CRR, since an outdated calibration may no longer reflect market conditions or adequately represent the exposure profile.

For internal risk management, the calibration frequency also affects the quality of exposure numbers used in the institution's day-to-day risk management process. Frequent calibration ensures that risk measures remain aligned with current market dynamics, supporting effective decision-making and risk mitigation strategies.

Differences have been observed in the calibration frequency for volatility and correlation parameters within risk factor simulation processes. While half of the institutions perform monthly or more frequent calibrations, the remainder calibrate less often. Adopting regular calibration of underlying stochastic processes—such as drift, volatility, and correlation—is considered good practice. This approach minimizes the risk of noncompliance with regulatory requirements and ensures that exposures are accurately assessed.

In summary, calibration is a vital component of credit risk modeling. It ensures that models remain accurate, reliable, and reflective of current market conditions. By maintaining appropriately calibrated models, financial institutions can better manage credit risk exposures, comply with regulatory mandates, and enhance their overall risk management practices.

4.2 Key Tests

In the calibration of Probability of Default (PD) estimates, it is essential for institutions to perform key tests that ensure the accuracy and reliability of their credit risk models. Given the variability in practices regarding calibration at different levels—portfolio level versus grade or pool level—the guidelines specify particular requirements to enhance the robustness of PD estimates.

Calibration Level Considerations:

Institutions may choose to perform calibration at the portfolio level or at the grade or pool level. To account for these differing practices, the guidelines (specifically in paragraph 92) stipulate that:

- If calibration is performed at the *grade or pool level*, institutions should provide additional calibration tests at the level of the relevant calibration segment (which corresponds to the portfolio level if there is only one calibration segment).
- Conversely, if calibration is performed at the *portfolio level*, institutions should perform additional calibration tests at the level of the grade or pool.

This approach ensures that PD estimates are appropriate and representative across both granular and aggregated levels, enhancing the reliability of risk quantification.

Calibration Frequency:

The frequency with which calibration is performed is critical for maintaining accurate PD estimates over time. Calibration frequency is relevant for:

- *Regulatory Reporting*: Regular calibration is necessary to provide accurate and up-to-date PD estimates for compliance with regulatory requirements.
- *Internal Risk Management*: Frequent calibration supports the use test requirements set by Article 289 of the Capital Requirements Regulation (CRR), impacting activities such as line consumption and other risk management practices.

Assessment of Calibration Methods:

Institutions are required to assess the potential effects of their chosen calibration methods on the behavior of PD estimates over time. This includes:

- Ensuring that PD estimates at the grade level are representative of the long-run average default rates.
- Understanding the cyclicality introduced by different calibration approaches and mitigating undue volatility in capital requirements.

Clarity on Calibration Methods:

The guidelines provide clarity on the various calibration methods that are permissible under the CRR. Definitions and specifications include:

- *Definition of Calibration*: Calibration is distinguished from model development and is defined as the process that leads to appropriate risk quantification.
- *Alignment with Long-Run Average Default Rates*: Calibration ensures that PD estimates in a calibration sample correspond to the long-run average default rate at the level relevant for the applied calibration method.

By adhering to these key tests and considerations, institutions can enhance the accuracy and reliability of their PD estimates, ensuring compliance with regulatory requirements and supporting sound risk management practices.

Calibration Plots Calibration plots are essential tools for assessing the accuracy of probability estimates in risk models. They compare the predicted probabilities (e.g., Probability of Default or PD estimates) with the observed default rates, helping to evaluate how well a model is calibrated.

To create a calibration plot, follow these steps:

1. **Group the Data**: Divide the dataset into groups (e.g., risk grades or probability bins) based on the predicted probabilities.
2. **Compute Predicted Probabilities**: For each group, calculate the average predicted probability.
3. **Compute Observed Default Rates**: For each group, calculate the observed default rate over a specified time horizon.
4. **Plot the Results**: Create a plot with the average predicted probabilities on one axis and the observed default rates on the other axis.

An ideal calibration plot would show points lying close to the diagonal line where the predicted probabilities equal the observed default rates, indicating that the model's estimates are accurate on average.

Example: Creating a Calibration Plot in Python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Sample data: predicted probabilities and actual outcomes
# Replace the 'predicted_probabilities' and 'actual_outcomes' with your
    data
predicted_probabilities = np.array([...])  # Predicted PDs
actual_outcomes = np.array([...])          # Actual defaults (1 for
   default, 0 for non-default)

# Create a DataFrame
data = pd.DataFrame({
    'Predicted_PD': predicted_probabilities,
    'Actual_Default': actual_outcomes
})

# Define the number of bins
num_bins = 10

# Assign observations to bins based on predicted probabilities
data['Probability_Bin'] = pd.qcut(data['Predicted_PD'], q=num_bins,
   duplicates='drop')

# Calculate average predicted PD and observed default rate for each bin
calibration_data = data.groupby('Probability_Bin').agg({
    'Predicted_PD': 'mean',
    'Actual_Default': 'mean'
}).reset_index()
```

```
# Plot the calibration plot
plt.figure(figsize=(8, 6))
plt.plot(calibration_data['Predicted_PD'], calibration_data['
    Actual_Default'], marker='o', linestyle='', label='Observed Default
    Rate')
plt.plot([0, 1], [0, 1], linestyle='--', color='grey', label='Perfect
    Calibration')
plt.xlabel('Average Predicted Probability')
plt.ylabel('Observed Default Rate')
plt.title('Calibration Plot')
plt.legend()
plt.tight_layout()
plt.show()
```

In this code:

- We create a DataFrame with the predicted probabilities and actual outcomes. - The data is divided into bins using the 'pd.qcut' function, which creates quantile-based bins. - For each bin, we calculate the average predicted probability and the observed default rate. - We plot these values against each other and include the diagonal line representing perfect calibration.

Interpreting Calibration Plots

When interpreting calibration plots, consider the following:

- **Good Calibration**: Points lying close to the diagonal line indicate that the model's predicted probabilities match the observed default rates, suggesting good calibration.
- **Overestimation of Risk**: Points below the diagonal line indicate that the model is overestimating the risk (the predicted probabilities are higher than the observed default rates).
- **Underestimation of Risk**: Points above the diagonal line suggest that the model is underestimating the risk (the predicted probabilities are lower than the observed default rates).
- **Segment-Specific Performance**: Deviations in specific bins may highlight areas where the model performs poorly, indicating the need for model adjustments for certain risk segments.

Factors Affecting Calibration

Several factors can influence the calibration of a risk model:

- **Time Slices Used**: The number of time periods included in the calibration sample affects its characteristics. Using all available time points makes the sample representative of the variability over time, while using a single point focuses on comparability with the current portfolio.
- **Calibration Sample Timing**: The timing of information used for grade calibration (e.g., type 1 or type 3 calibration) impacts the calibration sample. It determines

whether the estimates are based on the most recent data or averaged over multiple periods.

- **Level of Calibration**: Calibration can be performed at different levels, such as portfolio level or grade/pool level. Each approach has implications for the cyclicality of capital requirements and the sensitivity of PD estimates to changes in the portfolio.

It's important to assess how the chosen calibration method affects the behavior of PD estimates over time. Performing additional calibration tests at various levels (e.g., at the portfolio level if calibration is done at the grade level, and vice versa) can provide insights into the robustness and stability of the model.

Best Practices

When working with calibration plots:

- **Use Sufficient Data**: Ensure that each bin has enough observations to produce reliable estimates of the observed default rates.
- **Assess Different Time Horizons**: Evaluate the calibration over different time periods to understand the model's performance under varying economic conditions.
- **Document Calibration Methods**: Clearly document the calibration methods used, including the level of calibration and the characteristics of the calibration sample.
- **Continuous Monitoring**: Regularly update and monitor calibration plots to detect any deterioration in model performance promptly.

By carefully creating and interpreting calibration plots, institutions can enhance their risk models, leading to more accurate PD estimates and better risk management practices.

4.2.1 Brier Score

The Brier Score is a key metric used to evaluate the accuracy of probability forecasts in binary outcomes, such as defaults in credit risk modeling. It measures the mean squared difference between the predicted probabilities and the actual outcomes, providing insight into the calibration of Probability of Default (PD) models.

Calculation of the Brier Score

To calculate the Brier Score, follow these steps for each observation in the dataset:

- *Predicted Probability*: Obtain the PD estimate for each individual borrower or exposure.
- *Actual Outcome*: Determine the actual default status, assigning a value of 1 if a default occurred and 0 otherwise.

- *Difference*: Subtract the predicted probability from the actual outcome.
- *Square the Difference*: Square the result to eliminate negative values and emphasize larger errors.

After performing these steps for all observations, compute the average of the squared differences. This average represents the Brier Score for the PD model under evaluation.

Interpretation of the Brier Score

The Brier Score ranges from 0 to 1, where:

- A score of **0** indicates perfect accuracy, meaning the predicted probabilities exactly match the actual outcomes.
- A higher score indicates poorer model performance, with **1** being the worst possible score.

In practice, a lower Brier Score signifies better model calibration, as it reflects smaller discrepancies between predicted PDs and observed default rates (DRs). This is crucial for financial institutions, as regulatory guidelines require accurate estimation of credit risk parameters.

Application in Model Validation

Regular calculation of the Brier Score is considered a best practice in validating PD models. It assists institutions in:

- *Back-testing PD Estimates*: By comparing PD best estimates without conservative adjustments to actual default occurrences, institutions assess the accuracy of their models.
- *Assessing Model Calibration*: The Brier Score highlights how well the predicted probabilities align with observed outcomes across different grades or pools.
- *Identifying Non-Monotonicity*: Analyzing variations in the Brier Score across rating grades can reveal inconsistencies in model performance, prompting further investigation.

Regulatory Considerations

Regulatory frameworks emphasize the importance of validating the predictive accuracy of PD models. The Brier Score serves as a quantitative tool in this context, aligning with the expectations outlined in regulatory guidelines. Institutions are encouraged to:

- Understand and regularly assess the Brier Score as part of their validation processes.
- Justify the approaches used in calculating observed average one-year default rates, such as choosing between overlapping or non-overlapping time windows.

- Analyze the drivers behind the Brier Score results to enhance model robustness and compliance.

Conclusion

In summary, the Brier Score is an essential metric for evaluating the accuracy of PD models in credit risk management. By providing a quantitative measure of the difference between predicted probabilities and actual outcomes, it aids institutions in ensuring their models are well-calibrated and comply with regulatory standards.

Hosmer-Lemeshow Test The Hosmer-Lemeshow test is a statistical method used to assess the goodness-of-fit for logistic regression models, particularly in the context of Loss Given Default (LGD) modeling. This test evaluates how well the estimated LGD values align with the realized LGD values across different segments of the data.

In this case, the test is applied to 12 LGD segments. A two-way contingency table is constructed with the dimensions of 12 rows and 12 columns. The rows represent the discretized estimated LGD values (LGD_{E_i}), while the columns represent the realized LGD values (LGD_{R_i}). Each cell in the table contains the observed frequency of defaulted facilities that correspond to the specific combination of estimated and realized LGD segments.

The process involves the following steps:

- **Discretization of LGD Values**: Both estimated and realized LGD values are discretized into 12 segments. This helps in categorizing the continuous LGD values into manageable groups.
- **Construction of the Contingency Table**: A 12x12 table is created to display all possible combinations of the discretized estimated and realized LGD segments.
- **Calculation of Observed Frequencies**: For each pair of LGD segments, the number of defaulted facilities falling into that combination is counted.
- **Assessment of Fit**: The observed frequencies are compared to the expected frequencies under the assumption that the model fits well. Significant deviations may suggest discrepancies between the estimated and realized LGD values.

By analyzing the differences between the observed and expected frequencies, the Hosmer-Lemeshow test provides insight into the model's predictive accuracy. A good fit indicates that the model's estimated LGD values closely match the realized LGD values across the different segments, while a poor fit suggests the need for model refinement.

This approach is valuable in back-testing, ensuring that the LGD estimates used for regulatory compliance are robust and reliable. It complements other statistical tests, such as the one-sample t-test for paired observations, by offering a non-parametric method to evaluate model performance without assuming a specific distribution of residuals.

Binomial Test The Binomial Test is a statistical hypothesis test used to determine whether the observed proportion of successes in a binary outcome experiment differs significantly from a specified theoretical proportion. In the context of model validation in finance, it is often employed to assess whether the frequency of an event (such as default or failure) in a sample aligns with the expected probability predicted by a model.

In this test, each trial is independent and has only two possible outcomes: success or failure. The test evaluates the null hypothesis that the true probability of success equals a specific value against the alternative hypothesis that it does not equal that value. By analyzing the number of observed successes in the sample, the Binomial Test calculates the probability of observing such results under the null hypothesis.

The p-value is a crucial component of the Binomial Test. It represents the probability of obtaining a result as extreme as, or more extreme than, the observed outcome if the null hypothesis is true. A low p-value indicates that the observed proportion is unlikely under the null hypothesis, suggesting a significant difference between the observed and expected proportions.

In financial model validation, the Binomial Test helps determine if discrepancies between predicted and actual outcomes are due to random chance or indicative of model deficiencies. For instance, if a credit risk model predicts a default probability of 2%, but the observed default rate is significantly higher, the Binomial Test can assess whether this difference is statistically significant.

The test is particularly useful when dealing with small sample sizes or when the normal approximation is not suitable. For larger samples, the test statistic may approximate a normal distribution, allowing for simplified calculations of the p-value through standard normal distribution functions. As noted, the test statistic is asymptotically normal, and the p-value used for the assessment can be calculated using the cumulative distribution function of the standard normal distribution.

For further details on interval estimation for a binomial proportion, see Brown, Cai, and DasGupta (2001)[1].Jeffreys Test

The Jeffreys test is conducted to assess the accuracy and calibration of the estimated Probability of Default (PD) for individual rating grades. This test is based on the number of customers (N), as defined in point (<>)(g) of Section (<>)2.5.1. The relevant PDs are those assigned to non-defaulted customers at the beginning of the relevant observation period.

For each rating grade, as well as for the overall portfolio, the Jeffreys test compares the observed default rates with the estimated PDs. The number of tests to be conducted corresponds to the number of rating grades for non-defaulted exposures, plus one additional test for the overall portfolio.

The input parameters required for the Jeffreys test include:

- **Name of the rating grade**: Identification of each rating grade under consideration.

[1]Brown, L., Cai, T., and DasGupta, A. (2001). "Interval Estimation for a Binomial Proportion." *Statistical Science*, Vol. 16(2), pp. 101–117.

- **Number of customers (N)**: The total number of non-defaulted customers in each rating grade at the beginning of the observation period.
- **Estimated PD**: The PD assigned to each rating grade.
- **Observed defaults**: The number of customers who defaulted during the observation period.

The results of the Jeffreys test involve calculating confidence intervals for the true default rates using a Bayesian approach, as outlined by Brown et al. (2001)[2]. This method accounts for the statistical uncertainty inherent in the observed default rates due to the finite sample size.

By comparing the estimated PDs with the confidence intervals derived from the observed default rates, institutions can assess whether the PD estimates are sufficiently accurate or if adjustments may be necessary. The Jeffreys test thus provides a rigorous statistical tool for model validation in the context of credit risk management.

All input parameters and test results are documented, with the data basis being the number of customers (N) at the beginning of the observation period, including the original exposures as specified[3].

4.3 Confidence Intervals

In this section, we explain how to calculate confidence intervals for the estimated Expected Exposure at the Portfolio Level (EEPE) when using Monte Carlo simulation methods with pseudo-random number generators. Confidence intervals provide a range within which the true EEPE is expected to lie with a certain level of confidence, typically 95

When estimating EEPE through simulations, it's crucial to assess the accuracy and reliability of the estimate. This is achieved by constructing confidence intervals that account for the inherent variability in the simulation process.

There are two methods to assess the confidence interval of the estimated EEPE:

1. **Method 1**:
   - *Assumption*: The EEPE follows a normal distribution.
   - *Interpretation*: The estimation error can be viewed as half the length of the 95% two-sided confidence interval centered around the estimated EEPE.
   - *Approach*:
     (a) Estimate the variance of the estimated EEPE using the results from the Monte Carlo simulations.
     (b) Calculate the standard error by taking the square root of the estimated variance.

[2]Brown, L., Cai, T., and DasGupta, A. (2001), "Interval Estimation for a Binomial Proportion," *Statistical Science*, Vol. 16(2), pp. 101–117.

[3]Refer to the original exposure 34 at the beginning of the relevant observation period.

   (c) Determine the confidence interval by multiplying the standard error by the appropriate critical value from the normal distribution (corresponding to the desired confidence level).

2. **Method 2**:

   - *Assumption*: Similar to Method 1, it assumes that the EEPE follows a normal distribution.
   - *Difference from Method 1*: It uses a different estimator for the variance of the estimated EEPE and does not apply a convergence adjustment.
   - *Interpretation*: The estimation error is again interpreted as half the length of the 95% two-sided confidence interval centered around the estimated EEPE.
   - *Approach*:
     (a) Approximate the variance of the estimated EEPE using an alternative estimation technique.
     (b) Calculate the standard error based on this variance estimate.
     (c) Determine the confidence interval using the standard error and the corresponding critical value from the normal distribution.

In both methods, the key steps involve estimating the variance of the estimated EEPE and using this to compute the standard error. The critical value from the normal distribution (e.g., approximately 1.96 for a 95% confidence level) is then used to calculate the width of the confidence interval.

It's important to note that the true variance of the estimated EEPE is often unknown and must be approximated. This is typically done using the sample variance obtained from the Monte Carlo simulations.

By constructing confidence intervals, practitioners can assess the precision of the estimated EEPE and understand the range of potential values it may take due to simulation variability. This is essential for risk management and regulatory compliance, as it provides insight into the reliability of the exposure estimates.

In summary, calculating confidence intervals for the estimated EEPE involves:

- Assuming a normal distribution for the EEPE.
- Estimating the variance of the EEPE from simulation results.
- Calculating the standard error as the square root of the estimated variance.
- Determining the confidence interval using the standard error and the appropriate critical value.

This process allows for a quantitative assessment of the uncertainty around the EEPE estimate and supports better decision-making in financial risk management.

Low Default Portfolio Tests Low Default Portfolios (LDPs) are characterized by a scarcity of observed defaults, which poses significant challenges in estimating risk parameters such as Probability of Default (PD) and Loss Given Default (LGD). These portfolios often include exposures to sovereigns, banks, and large corporates where default events are rare but can have substantial impact when they occur.

An appropriate solution for the definition and treatment of LDPs involves several key considerations:

- **Clear Definition**: Establish precise criteria to identify LDPs based on historical default rates and the nature of the exposures. This ensures consistency in how portfolios are classified and managed.
- **Modelling Restrictions**: Due to the limited default data, traditional statistical models may not be suitable. Restrictions may include:
  - Using conservative estimates and adding margins of conservatism to account for uncertainty.
  - Incorporating external data sources or benchmarks to supplement internal data.
  - Applying expert judgment carefully documented and subject to rigorous oversight.
- **Regulatory Compliance**: Aligning with regulatory expectations, such as those outlined by the Basel Committee on Banking Supervision (BCBS)[4], to ensure that the models used are appropriate for LDPs and that the capital requirements reflect the underlying risk.
- **Validation Practices**: Implementing robust validation techniques tailored to LDPs, acknowledging the limitations posed by low default rates. This may involve stress testing and scenario analysis.

It is important to note that during the Targeted Review of Internal Models (TRIM), the term *low-default portfolio* was used for simplicity. As indicated in Section 3.2.3, some models assigned to the LDP category were not strictly LDP models, while others in the *retail and SME* category exhibited characteristics of LDPs[5]. This highlights the necessity for precise definitions and careful treatment of such portfolios.

In conclusion, addressing the challenges of LDPs requires a combination of clear definitions, appropriate modelling restrictions, adherence to regulatory guidance, and specialized validation practices. By doing so, financial institutions can better manage the risks associated with low default portfolios and ensure that their risk assessments and capital allocations accurately reflect the underlying credit risk.

[4] See http://www.bis.org/bcbs/publ/d298.pdf.

[5] As indicated in Section 3.2.3 of the TRIM guidelines.

4.4 Benchmarking Calibration

Benchmarking calibration is a critical component in the validation of risk models, ensuring that the estimated parameters accurately reflect the true risk profile of the institution. This process involves comparing the model's outputs against both internal and external benchmarks to assess performance and identify potential areas of improvement.

A primary consideration in benchmarking is the selection of appropriate calibration methodologies based on the quality and quantity of available data. When sufficient internal data is present, the calibration should primarily rely on this data to capture the specific characteristics of the institution's exposures. The validation function is expected to verify that the reliance on internal data is justified and that the chosen calibration approach is suitable.

In situations where internal data is insufficient—such as with new portfolios, low default portfolios, or during periods of significant change—external data may be utilized to supplement the calibration. The use of external data should be carefully justified, ensuring that it is only used to enhance internal experience, not replace it. The validation function should confirm that this reliance on external data is a necessary response to data scarcity and that the external data is relevant and comparable.

When external benchmarks are available, such as industry averages or external credit ratings, they should be used as challengers rather than definitive standards. Comparing internal model outputs with external benchmarks can highlight discrepancies and potential weaknesses, prompting further investigation. However, these external benchmarks should not be the sole measure of model performance; they are tools to support a comprehensive validation process.

It is essential that the calibration process maintains the rank ordering of obligors or exposures consistent with the institution's risk assessment framework. Adjustments made for specific areas of use—such as adopting a lifetime horizon for Probability of Default (PD) estimates in compliance with accounting standards—should not alter the relative risk rankings established by the model. Altering the rank ordering could compromise the integrity of the model and fail to meet regulatory use test requirements.

Moreover, the validation function should ensure that there is a sufficient number of observed defaults and recovery processes to produce robust Loss Given Default (LGD) estimates. In cases of data scarcity, alternative methods such as pooling data from similar portfolios or extending observation periods may be necessary to enhance the reliability of the estimates.

In summary, benchmarking calibration against internal and external benchmarks is a nuanced process that requires:

- A thorough assessment of the availability and adequacy of internal data.
- Justification for the use of external data when internal data is lacking.
- Use of external benchmarks as tools for challenging and improving the model, not as absolute standards.

- Maintenance of consistent rank ordering of exposures to preserve the model's integrity.
- Verification that sufficient data supports the robustness of parameter estimates.

By meticulously addressing these considerations, institutions can enhance the accuracy and reliability of their risk models, ensuring they are well-calibrated to reflect both internal experiences and relevant external information.

5 PD Stability Tests

Assessing the stability of Probability of Default (PD) models over time is crucial to ensure their reliability and robustness. Stability tests help validate that the PD estimates remain consistent across different time periods and market conditions, which is essential for accurate risk assessment and regulatory compliance.

Stability testing verifies that the PD model's performance is not unduly influenced by specific time periods or data samples. A stable PD model should provide consistent predictive power, regardless of temporal fluctuations or changes in the economic environment.

Common methods used to assess the stability of PD models include:

- **Time Slice Analysis**: Evaluating model performance across multiple time periods, or *time slices*, to detect any temporal inconsistencies. Using all time slices contained in the development sample was the most common practice, applied in 48% of all PD models. However, a significant number of models (23%) are calibrated using only one time slice. Institutions that consider all time slices may also integrate external data to enhance the analysis.
- **Population Stability Index (PSI)**: Measuring changes in the distribution of input variables or model scores over time to identify shifts that may affect model performance.
- **Characteristic Stability Analysis**: Comparing the stability of individual model characteristics or risk drivers to ensure they remain predictive over time.

An effective analysis of PD model stability should encompass the following characteristics:

1. **Comprehensive Performance Evaluation**: Assessing the model's accuracy and discriminatory power across different time periods to ensure consistent performance.
2. **Identification of Model Changes**: Monitoring for significant shifts in model parameters or performance metrics that may indicate the need for model recalibration or redevelopment.

The type of PD model employed can influence stability testing approaches. According to the IRB survey, the majority of PD models (approximately 90%) are based on scorecards—either utilizing expert judgment or quantitative data. Scorecards based on quantitative data account for 63% of all PD models and 65% of all exposures covered by PD models in the sample. For retail exposure classes, such quantitative scorecard models are used exclusively.

Understanding the model type helps in tailoring stability tests appropriately. For instance, models based on quantitative data may require rigorous statistical analysis to detect subtle changes in predictive variables, whereas models relying on expert judgment may focus more on qualitative assessments.

Regulatory expectations emphasize regular stability testing of PD models to ensure ongoing compliance and robustness. The analysis should be thoroughly documented, and any identified issues should be addressed promptly. The appendix of relevant regulatory reports often outlines how different scenarios are classified to assess whether a model change is necessary.

In conclusion, PD stability tests are essential for maintaining the integrity of credit risk models. By systematically evaluating model performance over time and under various conditions, institutions can ensure their PD models remain reliable tools for risk assessment and decision-making.

5.1 Introduction to Model Stability

Model stability is a critical aspect of credit risk modeling, referring to the consistency and reliability of a model's performance over time. It encompasses the ability of a model to produce accurate and robust estimates of default probabilities across different economic conditions and time horizons. Ensuring model stability is essential for financial institutions to maintain confidence in their risk assessments and to meet regulatory requirements effectively.

In the context of credit risk models, particularly those dealing with low default portfolios (LDPs), data availability and representativeness significantly influence model stability. Institutions often rely on historical default data to define the likely range of variability of one-year default rates. However, in many cases, the period considered is constrained by the availability of data, typically starting from the mid-2000s corresponding to the inception of Basel II standards. This limitation can hinder the model's ability to capture a full spectrum of economic cycles, including periods of financial stress.

A noteworthy observation is that several models have not been recalibrated after significant market events, such as the financial crisis of 2008-09. The lack of model recalibration over such extended periods is questionable unless justified by thorough validation of the estimates. Regular recalibration is paramount to ensure that models remain accurate and reflective of the current risk environment.

Another crucial factor impacting model stability is the time horizon considered in the risk assessment. According to the European Central Bank (ECB) guide, the rating or grading assignment process should not only focus on a one-year horizon but also adequately anticipate and reflect risk over a longer time frame. This approach involves taking into account plausible changes in economic conditions and balancing predictive drivers that are effective over both short-term and forward-looking horizons.

Despite this guidance, a horizontal analysis revealed that in 41% of cases, institutions adopted a time horizon of one year, and 24% did not explicitly define a time horizon. Only 35% of institutions used a time horizon greater than one year. This tendency towards short-term horizons can limit the model's ability to anticipate future risks and may compromise its stability over time.

Variations in practices across different jurisdictions further impact model stability. In countries without explicit regulatory guidance on the time series to be used, institutions tend to rely solely on available observed data since the implementation of Basel II. Conversely, in countries with explicit guidance, institutions may incorporate more extended historical data, enhancing the model's ability to account for various economic scenarios.

To improve model stability, institutions should consider the following:

- **Regular Model Recalibration:** Continuously update models to incorporate recent data and reflect current market conditions.
- **Extended Time Horizons:** Adopt a longer time horizon in the rating assignment process to capture long-term risks and anticipate future changes.
- **Balanced Predictive Drivers:** Ensure an appropriate balance between short-term predictive factors and forward-looking indicators.
- **Consideration of Economic Conditions:** Account for plausible changes in economic conditions and stress scenarios in the modeling process.
- **Adherence to Regulatory Guidance:** Follow the ECB guide and other regulatory standards to align practices with industry best practices.

By focusing on these areas, institutions can enhance the stability of their credit risk models. Stable models contribute to more reliable risk assessments, better capital allocation, and overall improved resilience of the financial system.

5.2 Key Tests

To evaluate the stability of internal ratings and risk parameters over time, several key tests are performed. These tests analyze the consistency and robustness of ratings assigned to obligors or facilities during a specific observation period. The primary tests for assessing stability include:

- **Migration Analysis**: This involves examining the transitions of ratings assigned to individual obligors or facilities from one period to another using a migration matrix. By tracking the movement between rating grades 1 to K at the beginning and end of the observation period, institutions can assess how stable the ratings are over time.
- **Alignment with Rating Philosophy**: The outcomes of the migration analysis are compared to the expected results based on the institution's rating philosophy. This helps in understanding the core features of the model concerning its rating philosophy. Significant deviations from expectations may indicate deficiencies in the model, such as missing risk drivers or inadequately defined grades or pools, leading to a lack of homogeneity.
- **Validation of Model Design Stability**: The validation function reviews the rating philosophy and stability properties of the model to ensure they are appropriate for the intended scope of application. This includes assessing the impact of risk quantification methodologies on the stability of risk parameters and determining whether the model's design remains effective over time.

- **Incorporation into Back-Testing**: The results of the stability analyses are utilized in back-testing procedures to evaluate the predictive accuracy of the risk parameters. By considering the stability findings, institutions can enhance the reliability of their back-testing and make informed decisions about model adjustments.

Performing these tests allows institutions to identify and address potential issues in their internal rating systems. By ensuring the stability of ratings and risk parameters, they can maintain robust risk management practices and comply with regulatory expectations.

Population Stability Index (PSI) The Population Stability Index (PSI) is a statistical measure used to assess the shift in the distribution of a variable between two different populations or time periods. In the context of finance and regulatory compliance, PSI is instrumental in monitoring the stability of risk models by detecting changes in the characteristics of the populations on which these models are applied.

Calculation of PSI

To calculate the PSI, follow these steps:

1. *Divide the data into bins*: Segment both the initial (baseline) population and the current population into the same set of categories or intervals based on the variable of interest.
2. *Calculate proportions*: For each bin, determine the proportion of observations in both the initial and current populations.
3. *Compute the PSI contribution for each bin*: For each bin, calculate the difference between the current and initial proportions, then multiply this difference by the logarithm of the ratio of the current proportion to the initial proportion.
4. *Sum the contributions*: Add up the PSI contributions from all bins to obtain the overall PSI value.

Interpretation of PSI

The PSI value indicates the degree of change in the population's characteristics:

- **PSI $<$ 0.1**: No significant change. The population is stable, and the model remains reliable.
- **PSI between 0.1 and 0.25**: Moderate change. The population has shifted, and the model may require review or recalibration.
- **PSI $>$ 0.25**: Significant change. The population has changed substantially, and the model may no longer be valid without adjustments.

Regular monitoring of PSI helps institutions ensure that their models accurately reflect current conditions and remain compliant with regulatory standards.

Considerations in PSI Assessment

When evaluating PSI, consider the following:

- *Sample Size*: Ensure that both populations have sufficient data to produce reliable proportions in each bin.
- *Bin Selection*: Use meaningful and consistent bins across populations to accurately capture distribution shifts.
- *External Factors*: Acknowledge that changes in economic conditions, such as macroeconomic shifts, can influence PSI values.

Statistical Significance and P-Values

In statistical analysis, understanding the significance of observed changes is crucial. P-values are used to determine the probability that the observed change could have occurred by chance. When the test statistic follows a normal distribution asymptotically, the p-value is calculated using the cumulative distribution function of the standard normal distribution evaluated at the test statistic.

For example, in assessing the stability of transition probabilities between different states (denoted by i and j), the p-value can be obtained by evaluating the cumulative distribution function at the test statistic $z_{i,j}$. A lower p-value indicates that the observed change is statistically significant.

Reporting Missing Values

In situations where the test statistic is not well-defined—perhaps due to zero frequencies or insufficient sample sizes—the corresponding z-values and p-values cannot be calculated. In such cases, these values should be reported as missing in the reporting templates to maintain transparency and accuracy in the analysis.

Impact of Macroeconomic Conditions

Institutions employing models sensitive to macroeconomic conditions, such as those following a point-in-time (PIT) rating philosophy, may experience variations in PSI that reflect economic cycles. Conversely, models aiming for through-the-cycle (TTC) stability might show less fluctuation. The calibration approach significantly influences how sensitive probability estimates are to economic environments:

- *PIT Models*: More responsive to current economic conditions, possibly resulting in higher PSI values during economic shifts.
- *TTC Models*: Aim for stability over time, potentially exhibiting lower PSI values despite economic changes.

Understanding these dynamics is essential for interpreting PSI values accurately and determining whether changes warrant model adjustments or are expected given the economic context.

5.2.1 Characteristic Stability Index (CSI)

The Characteristic Stability Index (CSI) is a statistical measure used to assess the stability of a specific characteristic or variable within a population over time. It is instrumental in monitoring shifts in the distribution of key variables, which can have significant implications for predictive models and regulatory compliance in the financial industry.

Calculation of CSI To calculate the CSI between a base period (e.g., model development period) and a comparison period (e.g., current portfolio), follow these steps:

1. **Segment the Data**: Divide the range of the characteristic into bins or categories. These bins should be mutually exclusive and collectively exhaustive, ensuring that each observation falls into one bin only.

2. **Calculate Proportions**: For each bin, compute the proportion of observations in both the base period and the comparison period. This involves counting the number of observations in each bin for both periods and dividing by the total number of observations in each period.

3. **Compute the CSI**: For each bin, calculate the difference between the base period proportion and the comparison period proportion. Multiply this difference by the natural logarithm of the ratio of the base period proportion to the comparison period proportion. Sum these values across all bins to obtain the CSI.

Interpretation of CSI The CSI quantifies the degree of change in the distribution of a characteristic over time. The interpretation of the CSI values is as follows:

- **CSI less than 0.1**: Indicates that the characteristic is stable, and no significant change has occurred.

- **CSI between 0.1 and 0.25**: Suggests a moderate shift in the characteristic's distribution, which may require monitoring.

- **CSI greater than 0.25**: Reflects a significant shift, signaling potential issues that could affect model performance and may necessitate further investigation.

Application in Model Validation In the context of model validation, the CSI is a crucial tool for detecting shifts in the variables used by credit models. Regular calculation of the CSI helps institutions ensure that the characteristics upon which models are built remain valid over time. For example, significant changes in customer behavior or economic conditions may alter the distribution of key variables, impacting the predictive power of models.

Considerations for Effective Use When utilizing the CSI, consider the following:

- **Bin Selection**: The choice of bins can significantly influence the CSI. Bins should be defined in a way that captures meaningful differences in the data.

- **Data Quality**: Accurate and complete data are essential for reliable CSI calculations. Ensure that data collection processes are robust.

- **Regular Monitoring**: Periodic calculation of the CSI allows for timely detection of shifts, enabling proactive management of model risk.
- **Sample Size**: Adequate sample sizes in both the base and comparison periods are necessary to derive meaningful conclusions from the CSI.

Relation to Population Stability Index (PSI) While the CSI focuses on the stability of a single characteristic, it is closely related to the Population Stability Index (PSI), which assesses the stability of the overall population or portfolio. Both indices are valuable in monitoring changes that can affect credit risk models and ensuring compliance with regulatory expectations for model performance monitoring.

Regulatory Context Regulatory bodies emphasize the importance of monitoring model inputs and performance. The use of indices like the CSI aligns with regulatory requirements for ongoing model validation and risk management. For instance, institutions may be required to report on measures like the Population Stability Index for facilities covered by a Credit Conversion Factor (CCF) approach, comparing the number of facilities at the beginning and end of an observation period.

Conclusion The Characteristic Stability Index is a vital tool for financial institutions to monitor and manage changes in key variables over time. By understanding and applying the CSI, institutions can maintain the integrity of their predictive models, comply with regulatory standards, and proactively address potential risks arising from shifts in characteristic distributions.

5.2.2 Divergence Measures

In the context of model validation for regulatory compliance in finance, divergence measures are essential tools used to quantify the differences between estimated probabilities and observed default rates (DRs). These measures help in assessing the performance of Probability of Default (PD) models by comparing the predicted risks with actual outcomes.

Understanding and correctly calculating divergence measures allows financial institutions to identify potential biases, model weaknesses, and areas for improvement in their credit risk assessment models.

Calculating Divergence Measures

To calculate divergence measures, one needs to compare the PD estimates assigned by the model to different segments (grades or pools) of the portfolio with the actual default rates observed in those segments over time. This involves:

1. *Data Collection*: Gather historical data on PD estimates and observed DRs for each grade or pool. Ensure that the data covers an adequate time horizon to capture different economic cycles, avoiding survivor bias, as methods should be carefully assessed to ensure they do not contain such bias.

2. *Model Application*: Apply the PD model backward in time to estimate the long-run average DR per grade or pool. As reported, a majority (65%) of institutions using type 1 and 3 calibrations follow this approach. Be aware that some institutions apply other methods, but details on these alternatives are often not disclosed.

3. *Comparison of Estimates and Observations*: For each grade or pool, compare the PD estimates with the actual DRs. This comparison can be performed using various statistical divergence measures.

Interpreting Divergence Measures

Interpreting divergence measures involves assessing the extent to which the PD estimates align with the observed DRs. A significant divergence may indicate that the model is not accurately capturing the risk and may require recalibration or refinement.

Key points to consider when interpreting divergence measures:

- *Consistency Over Time*: Assess whether the divergence is consistent across different time periods, especially during various economic conditions.
- *Impact of Qualitative Components*: Understand that incorporating qualitative components into the model can affect the divergence measures, although methods for applying qualitative factors backward in time are often not well-explained.
- *Regulatory Requirements*: Ensure compliance with regulatory guidelines, such as those specified in Article 180(1)(a) and Article 169(3) of the Capital Requirements Regulation (CRR), which outline the standards for PD estimations and calibrations.

Avoiding Survivor Bias

When calculating divergence measures, it is crucial to avoid survivor bias. Survivor bias occurs when only the entities that have survived until the end of the observation period are included in the analysis, leading to an overestimation of model performance. To mitigate this:

- Include all relevant data, not just data from entities that have not defaulted.
- Use methods that account for the full population of obligors over the entire observation period.
- Regularly assess and adjust the model to account for changes in the portfolio and external factors.

Example Implementation in Python

Below is a simplified example of how divergence measures can be calculated using Python. This code compares PD estimates with observed DRs for different grades.

```python
import pandas as pd

# Sample data: Replace with actual PD estimates and observed DRs
```

```
data = {
    'Grade': ['A', 'B', 'C', 'D'],
    'Estimated_PD': [0.01, 0.02, 0.05, 0.10],
    'Observed_DR': [0.015, 0.018, 0.055, 0.095]
}

df = pd.DataFrame(data)

# Calculate divergence as the difference between Estimated PD and
   Observed DR
df['Divergence'] = df['Estimated_PD'] - df['Observed_DR']

# Calculate absolute divergence
df['Absolute_Divergence'] = df['Divergence'].abs()

# Output the results
print("Divergence Measures per Grade:")
print(df[['Grade', 'Estimated_PD', 'Observed_DR', 'Divergence', '
   Absolute_Divergence']])
```

Interpreting the Results

The output of the code provides divergence measures per grade. A small divergence indicates good alignment between the model's PD estimates and the observed DRs. Larger divergences may signal the need for model adjustment. The absolute divergence helps in assessing the magnitude of the difference regardless of direction.

Conclusion

Calculating and interpreting divergence measures is a critical step in validating PD models. By carefully analyzing the differences between estimated probabilities and observed outcomes, financial institutions can enhance model accuracy, ensure compliance with regulatory standards, and ultimately improve risk management practices. It is essential to document the considerations for choosing specific approaches, as this transparency can lead to better regulatory compliance and internal understanding of the model's performance.

6 PD Backtesting

PD (Probability of Default) backtesting is a critical process in credit risk management that evaluates the accuracy and reliability of PD models used by financial institutions. It involves comparing the predicted probabilities of default with the actual observed default rates over a specific period. This comparison helps institutions assess the performance of their PD models and ensure compliance with regulatory requirements.

Regulatory back-testing is mandatory under the Capital Requirements Regulation (CRR) and has a direct impact on the amount of own funds requirements via the back-testing addend. According to Article 366(3) of the CRR, regulatory back-testing compares the hypothetical and actual changes in the portfolio's value ("hypothetical P&L" and "actual P&L") with the related one-day Value at Risk (VaR) number generated by the institution's model. Therefore, the changes in value of all, and only, the instruments and transactions included in the scope of calculation of the VaR model should be considered in the calculation of the hypothetical and actual P&L.

The European Central Bank (ECB) has provided guidance to clarify the back-testing framework in terms of scope, definitions, methodology for calculating actual and hypothetical P&Ls, and the counting and analysis of overshootings—cases where the P&Ls exceed the risk numbers. The ECB guide includes concrete examples of when overshooting notifications can be withdrawn, leveraging experiences collected through the Targeted Review of Internal Models (TRIM).

Institutions are expected to have a documented policy and procedure describing how they calculate the actual and hypothetical P&L, in accordance with Article 368(1)(e) of the CRR. The ECB considers that, to be fit for purpose, the policy and procedure should include at least the following key information:

- The methodology used for calculating actual and hypothetical P&L.
- The scope of instruments and transactions included in the P&L calculations.
- Procedures for identifying and analyzing overshootings.
- Criteria and processes for withdrawing overshooting notifications.
- Documentation of any assumptions or approximations used in the calculations.

For exposures to corporates, institutions, central governments, and central banks where the obligors are highly leveraged or have predominantly traded assets, competent authorities must verify that the PD reflects the performance of the underlying assets during periods of stressed volatility, as referred to in Article 180(1)(a) of the CRR.

To assess the appropriateness and implementation of the policy and procedure for calculating actual and hypothetical P&L, the ECB can, based on Article 10 of the Single Supervisory Mechanism (SSM) Regulation, require institutions to provide detailed decompositions of economic, actual, and hypothetical P&L into their elements for a sample of transactions or portfolios.

In practice, PD backtesting involves several key techniques:

- **Calibration Tests**: Assessing whether the predicted PDs accurately reflect observed default rates over time.
- **Discriminatory Power Analysis**: Evaluating the model's ability to distinguish between defaulting and non-defaulting obligors.
- **Stability Analysis**: Monitoring the stability and consistency of PD estimates over different periods.

By rigorously backtesting PD models, institutions can ensure that their models are accurate, reliable, and compliant with regulatory requirements. This process supports effective risk management and helps maintain the integrity of the financial system.

6.1 Introduction to Backtesting

Backtesting is a fundamental component of credit risk model validation. It involves comparing the predictions of a credit exposure model with actual observed outcomes to assess the model's accuracy and reliability. In the context of counterparty credit risk (CCR), backtesting serves to validate the effectiveness of exposure models in capturing the true risk associated with counterparties under various market conditions.

The primary purpose of backtesting in credit risk model validation is to ensure that the model's key assumptions and risk measures accurately reflect real-world outcomes. This process allows institutions to identify any poor performance in their exposure models and make necessary adjustments to improve their predictive capabilities.

According to Article 294(1)(h) of the Capital Requirements Regulation (CRR), backtesting samples must be representative and selected based on their sensitivity to material risk factors and their combinations. This means that the samples used should cover a wide range of scenarios that are relevant to the institution's risk profile. Furthermore, as stated in Article 294(1)(j) of the CRR, the backtesting program must be capable of identifying poor performance of an Expected Positive Exposure (EPE) model's risk measures.

To achieve a meaningful assessment of the CCR exposure model, institutions should:

- **Use Representative Backtesting Samples:** Select samples that are sensitive to key risk factors and their combinations to ensure comprehensive coverage of potential risk scenarios.
- **Conduct Backtesting at Multiple Levels:** Perform backtesting at both the risk factor level and the portfolio level (including actual and hypothetical portfolios) to capture all key assumptions of the CCR exposure model, especially for complex instruments like non-plain vanilla transactions or margined netting sets.
- **Calculate Backtesting Coverage Ratios:** Determine the shares of back-tested risk factors or portfolios to ensure that the backtesting framework provides sufficient coverage and identifies any gaps in the model's assessment.

- **Test Key Model Assumptions:** Design backtesting to evaluate the key assumptions, including the relationships between different tenors of the same risk factor and between different risk factors.
- **Perform Regular Backtesting:** Conduct backtesting at least once a year, as recommended under Article 293(1)(b) of the CRR, to maintain the model's integrity over time.

Incorporating backtesting into the model validation process helps institutions to:

- **Identify Model Weaknesses:** Detect areas where the model may not accurately predict exposures, allowing for timely improvements.
- **Ensure Regulatory Compliance:** Meet the requirements set forth in the CRR and other regulatory guidelines by demonstrating the model's reliability and robustness.
- **Enhance Risk Management:** Strengthen the overall risk management framework by ensuring that exposure measurements are accurate and reflective of actual risk.

In summary, backtesting is essential for verifying the predictive power of credit risk models and ensuring that they are fit for purpose. By rigorously testing models against historical data, institutions can enhance the credibility of their risk assessments and maintain confidence in their ability to manage counterparty credit risk effectively.

6.2 Backtesting Techniques

Regulatory back-testing is a mandatory requirement under the Capital Requirements Regulation (CRR) and has a direct impact on the amount of own funds requirements through the back-testing addend. The European Central Bank (ECB) has provided guidance to clarify its understanding of the back-testing framework, particularly in terms of scope, definitions used, methodologies for calculating actual and hypothetical Profit and Loss (P&L), and the counting and analysis of *overshootings*—instances where the P&Ls exceed the risk estimates.

The ECB guide leverages experience and examples collected through the Targeted Review of Internal Models (TRIM) to provide concrete scenarios where overshooting notifications could or could not be withdrawn. In cases of overshootings, institutions are expected to conduct detailed analyses to determine the underlying causes and assess the implications for their internal models.

Under Article 368(1)(e) of the CRR, institutions are required to have a documented policy and procedure detailing how they calculate the actual and hypothetical P&L. To be fit for purpose, this policy and procedure should include at least the following key information:

- The methodologies used for calculating actual and hypothetical P&L.
- Definitions and scope applied in the calculations.

- Procedures for identifying and analyzing overshootings.
- Governance processes surrounding back-testing activities.

Moreover, Article 369(2) of the CRR stipulates that back-testing performed in internal validation must comply with the same requirements as regulatory back-testing concerning the calculation of actual and hypothetical P&L. Therefore, the requirements described regarding the calculation of actual and hypothetical P&L should also be applied to internal back-testing to ensure consistency. When verifying compliance with this provision, the ECB will take into account the specific circumstances of the institution.

To assess the appropriateness and implementation of the policies and procedures for calculating the actual and hypothetical P&L, the ECB can, based on Article 10 of the Single Supervisory Mechanism (SSM) Regulation, require institutions to provide detailed decompositions of economic, actual, and hypothetical P&L into their elements for a sample of transactions or portfolios.

Traffic Light Approach

The Traffic Light Approach, as prescribed by the Basel requirements, is a common back-testing technique used to evaluate the performance of internal models. This approach categorizes institutions into different zones based on the number of overshootings observed:

- **Green Zone**: Indicates that the model is performing adequately, with an acceptable number of overshootings.
- **Yellow Zone**: Suggests potential issues with the model, warranting closer scrutiny and possibly leading to increased capital requirements.
- **Red Zone**: Reflects significant model deficiencies, leading to substantial increases in capital requirements and necessitating urgent corrective actions.

Institutions must monitor their back-testing results regularly and analyze any overshootings to understand their causes. The Traffic Light Approach provides a framework for regulators and institutions to determine the reliability of internal models and the adequacy of capital buffers.

In summary, robust back-testing techniques are essential for validating the accuracy of risk models and ensuring regulatory compliance. Institutions should maintain comprehensive policies and procedures for calculating actual and hypothetical P&L, as required by the CRR and clarified by the ECB. By adhering to these requirements and effectively utilizing approaches like the Traffic Light Approach, institutions can enhance their risk management practices and maintain confidence in their internal models.

Traffic Light Approach The Traffic Light Approach is a regulatory framework used to evaluate the performance of financial institutions' internal risk models, particularly in the context of market risk assessment and model validation. This approach categorizes the outcomes of backtesting results into different zones—**green**, **yellow**, and **red**—based on the number of exceptions observed, which are instances where actual losses exceed the model's predicted losses.

The primary objective of the Traffic Light Approach is to provide a standardized method for both institutions and competent authorities to assess the accuracy and reliability of risk models. This categorization aids in determining the adequacy of capital reserves and whether any adjustments or interventions are necessary to comply with regulatory requirements.

- **Green Zone**: Indicates that the model is performing within acceptable limits. The number of exceptions falls within the range expected due to random fluctuations. Institutions in this zone can continue using their models without additional regulatory capital requirements.
- **Yellow Zone**: Suggests that the model's performance is marginal and may not fully capture the risks. Institutions may be required to conduct further investigations into their models and could face increased capital requirements to buffer against potential underestimations of risk.
- **Red Zone**: Signals significant model deficiencies. The number of exceptions is unacceptably high, indicating that the model does not adequately capture the risk profile. Institutions are typically required to take immediate corrective actions, which may include model recalibration, redevelopment, or reverting to more conservative standardized approaches until issues are resolved.

In relation to the Capital Requirements Regulation (CRR), the Traffic Light Approach complements the regulatory framework by ensuring that internal models used for calculating risk-weighted exposure amounts are robust and reliable. Specifically, for equity exposures under the Internal Ratings-Based (IRB) Approach, Article 155 of the CRR outlines methods such as the simple risk weight approach, the PD/LGD approach, and the internal models approach. Institutions opting to use the internal models approach must meet additional requirements and obtain permission from competent authorities, as stipulated in Article 151(4) of the CRR.

The use of the Traffic Light Approach serves as an important tool for competent authorities when granting such permissions. It provides a clear and quantifiable measure of model performance, assisting in the assessment of whether an institution's internal models are compliant with regulatory standards. Furthermore, it ensures that institutions maintain sufficient capital buffers in alignment with their actual risk exposures, thereby promoting stability and resilience within the financial system.

In exceptional circumstances, such as the recent acquisition of new business, an institution may seek permission to use a combination of approaches (e.g., the Basic Indicator Approach and the Standardised Approach) during a transition period, as noted in Article 4 of the regulatory texts. The Traffic Light Approach can be instrumental during such periods by monitoring model performance and ensuring that risk assessments remain accurate throughout the transition.

Overall, the Traffic Light Approach is a vital component in the regulatory landscape, bridging the gap between model validation and regulatory compliance. It fosters an environment where institutions are encouraged to continually improve their risk assessment models, thereby enhancing the overall robustness of the financial sector.

Basel Backtesting Requirements In accordance with the Basel regulatory framework, financial institutions are required to perform back-testing on both actual and hypothetical changes in their portfolio's value. This process is essential to validate the accuracy of internal models used for market risk management and to ensure regulatory compliance.

Back-testing on *actual* changes involves comparing the portfolio's end-of-day value with its actual value at the end of the subsequent day, excluding fees, commissions, and net interest income. This method assesses the model's performance by reflecting real market movements and the impact of executed trades.

Conversely, back-testing on *hypothetical* changes is conducted by comparing the portfolio's end-of-day value with its value at the end of the subsequent day, assuming that all positions remain unchanged. This approach isolates the effect of market fluctuations on the portfolio's value, eliminating the influence of trading activities.

To evaluate the effectiveness of these back-testing procedures, the European Central Bank (ECB) may, under Article 10 of the Single Supervisory Mechanism (SSM) Regulation, request institutions to provide detailed decompositions of economic, actual, and hypothetical profit and loss (P&L) for selected transactions or portfolios. This enables the ECB to assess the appropriateness and implementation of the policies and procedures used in P&L calculations.

According to Article 294(1)(h) of the Capital Requirements Regulation (CRR), back-testing samples must be representative and selected based on their sensitivity to material risk factors and their combinations. Furthermore, as stated in point (j) of the same paragraph, an institution's back-testing program must be capable of identifying poor performance of an Expected Positive Exposure (EPE) model's risk measures.

As a result, the ECB expects back-testing samples to facilitate a meaningful assessment of the Counterparty Credit Risk (CCR) exposure model. Institutions should ensure comprehensive coverage within their back-testing framework by calculating back-testing coverage ratios—that is, the proportion of risk factors or portfolios that are back-tested—at least at the risk factor level and, if applicable, at the actual portfolio level. This comprehensive approach helps in:

- Identifying deficiencies in the risk measurement models.
- Ensuring that all material risk factors are adequately captured.
- Enhancing the robustness of the risk management framework.

By adhering to these requirements, institutions not only comply with regulatory standards but also strengthen their internal risk management practices.**Statistical Tests**

Statistical tests are essential tools in the validation of financial models, serving to rigorously assess model performance and robustness over time. The validation function is expected to form an opinion on the model's performance across different samples, particularly by comparing results from various time periods—such as conducting tests separately for each year of observation—and contrasting them with the performance achieved during model development.

Key aspects of employing statistical tests in model validation include:

- **Assessment Across Time**: Evaluating the model's predictive power over successive periods helps identify any deterioration or improvement in performance. This temporal analysis is crucial for understanding the model's stability and reliability.
- **Comparison with Development Phase**: By comparing current performance metrics with those from the model development phase, validators can ascertain whether the model continues to perform as expected or if recalibration is necessary.
- **Variance and Covariance Estimation**: For tests involving the Area Under the Curve (AUC), it's necessary to estimate the variance of the AUC or the covariance between AUCs from different rating assignments for a given portfolio during the observation period. These estimates are fundamental for conducting meaningful statistical comparisons.

Implementing appropriate statistical methodologies ensures that the validation process is thorough and complies with regulatory expectations. References to established research, such as the work by Brown and Benedetti[6], provide a solid foundation for selecting and justifying the use of specific tests.

By carefully choosing relevant statistical tests and performing comprehensive validation activities, organizations can maintain confidence in their models' predictive capabilities and uphold the standards required for regulatory compliance.

6.3 Backtesting Reporting

Backtesting reporting is a critical process in ensuring compliance with regulatory requirements, particularly under the Capital Requirements Regulation (CRR). It directly impacts the calculation of own funds requirements through the back-testing addend. The European Central Bank (ECB) has provided guidance to clarify the back-testing framework, including definitions, methodologies, and procedures for handling overshootings—instances where the Profit and Loss (P&L) exceeds the risk forecasts.

To effectively perform backtesting reporting, institutions should adhere to the following practices:

- **Daily Calculations on Global Business Days:** For every global business day, institutions must calculate the actual and hypothetical P&L, compute the Value at Risk (VaR), and conduct market risk monitoring and reporting. Calculations performed on non-global business days should *not* be used for regulatory backtesting purposes.

[6] Brown, M. and Benedetti, J. (1977), "Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables", *Journal of the American Statistical Association*, Vol. 72, pp. 309-315.

- **Consistent P&L Comparison:** The actual and hypothetical P&L used for backtesting should represent the P&L between two consecutive global business days. This P&L should be compared with the corresponding one-day VaR forecast for a one-day holding period between those two days, based on the portfolio composition at the start of the period.
- **Inclusion of All Positions:** All positions of trading units, including those in locations with local non-business days, must be included in the calculation of consolidated figures to ensure comprehensive reporting.
- **Overshootings Analysis:** In cases of overshootings—where the P&L exceeds the risk numbers—institutions must analyze and report these instances. The ECB provides concrete examples of when overshooting notifications can be withdrawn, leveraging experiences collected through the Targeted Review of Internal Models (TRIM).
- **Maintaining Time Series Data:** For top-of-the-house level reporting, institutions should maintain time series of p-values for daily actual P&L and hypothetical P&L against daily P&L forecasts from the VaR engine. This data should cover at least one year, preferably three years, to facilitate robust backtesting analysis.

The backtesting report should include:

- **Summary of Backtesting Results:** An overview of the backtesting outcomes, highlighting key findings and trends.
- **Detailed Overshootings Analysis:** Comprehensive analysis of any overshootings, including potential causes and remedial actions taken.
- **Comparative Charts and Tables:** Visual representations comparing actual and hypothetical P&L against VaR forecasts to illustrate performance.
- **Methodology Documentation:** Records of the methodologies used for calculations, including any changes or adjustments made during the reporting period.
- **Appendix with Time Series Data:** An appendix containing the complete time series of p-values and P&L data used in the analysis.

By following these guidelines, institutions can ensure their backtesting reporting is thorough, compliant with regulatory standards, and provides valuable insights into the effectiveness of their risk measurement models.

7 LGD Model Validation

The validation of Loss Given Default (LGD) models is essential to ensure the accuracy and reliability of credit risk assessments within financial institutions. The primary aim of the validation tools discussed in this section is to monitor the performance of LGD models, focusing particularly on their predictive ability or calibration. This involves evaluating how well the models predict actual losses when defaults occur.

7.1 Objectives of the Validation Tools

The validation tools are designed to assess LGD models' performance in several key areas:

- **Predictive Ability (Calibration):** Evaluating the accuracy of the LGD estimates produced by the models compared to observed losses.
- **Qualitative Analysis:** Ensuring the appropriate assignment of LGD estimates to the application portfolio and analyzing the distribution of LGD across different facilities under review.

7.2 Qualitative Aspects of LGD Models

Analyzing the qualitative aspects of LGD models involves a thorough examination of the methodologies and data used to assign LGD values. This ensures that the models are appropriately tailored to the specific characteristics of the portfolio and that the LGD distributions reflect the risk profiles of the facilities analyzed.

7.3 Definition of LGD Models

For the purposes of this validation, an LGD model is defined in line with regulatory guidelines, specifically referring to all data and methods used to assess the level of loss in the event of default for each facility covered by the model. This comprehensive definition ensures that all components contributing to the LGD estimation are included in the validation process.

7.4 Handling Missing or Adjusted LGD Estimates

In some cases, LGD model estimates may be unavailable for certain facilities due to a lack of data on critical risk drivers. When this occurs, predefined values or conservative fallback estimates are assigned. Additionally, institutions may impose caps or floors on LGD estimates, forcing them to adhere to specified thresholds. The validation process must account for these instances by:

- Identifying facilities where LGD estimates are missing or forced to default values.
- Assessing the impact of using predefined fallback values on the model's predictive ability.

- Evaluating the effects of applied caps or floors on the distribution and accuracy of LGD estimates.

By thoroughly examining these aspects, the validation ensures that the LGD models remain robust and that any adjustments or limitations are appropriately managed within the risk assessment framework.

7.5 Introduction to LGD Validation

Loss Given Default (LGD) is a critical component in credit risk modeling, representing the proportion of an exposure that is lost when a borrower defaults, after accounting for recoveries from collateral and other sources. Accurate estimation of LGD impacts the assessment of expected losses and capital requirements, making the validation of LGD models essential for financial institutions.

Validating LGD models poses specific challenges due to the inherent complexities of default and recovery processes. Unlike Probability of Default (PD) models, LGD models must consider a wide array of factors such as collateral valuation, recovery rates, legal costs, and the time value of money. Additionally, LGD estimation is affected by limited default data and prolonged recovery periods, which can complicate model calibration and performance monitoring.

One of the primary challenges in LGD validation is monitoring the model's predictive ability, or calibration. According to regulatory guidelines on PD and LGD estimation, institutions are expected to implement validation tools that rigorously assess the accuracy of LGD predictions. This involves not only statistical analysis but also an ongoing review of model assumptions and parameters to ensure they remain relevant under changing market conditions.

Furthermore, regulatory guidance on LGD modeling is extensive and continually evolving. Surveys have indicated that guidance affects numerous aspects of LGD models, often necessitating significant revisions. It is estimated that all LGD models may require changes in one or more dimensions to comply with new standards. This underscores the importance of a comprehensive validation process that can identify areas where models may not meet current regulatory expectations.

In practice, financial institutions must adopt a multidimensional approach to LGD validation. This includes:

- **Data Quality Assessment**: Evaluating the completeness and accuracy of data used in model development and validation.
- **Statistical Testing**: Applying rigorous statistical methods to assess model performance and predictive power.
- **Back-Testing**: Comparing model predictions with actual observed outcomes to measure accuracy.
- **Sensitivity Analysis**: Analyzing how changes in model inputs affect outputs to identify potential weaknesses.

- **Compliance Review**: Ensuring that models adhere to all relevant regulatory requirements and guidelines.

Effective LGD validation not only supports regulatory compliance but also enhances risk management practices by providing more reliable estimates of potential losses. As the regulatory landscape continues to evolve, ongoing validation efforts are crucial to maintain the integrity and effectiveness of LGD models.

7.6 Discriminatory Power Tests

The discriminatory power of Loss Given Default (LGD) models is crucial in ensuring that these models effectively distinguish between facilities with high and low LGD values. This capability is essential for accurate risk assessment and regulatory compliance in finance. An LGD model with good discriminatory power enables financial institutions to rank facilities according to their expected losses upon default, thus facilitating better decision-making in credit risk management.

Similar to the analyses conducted for Probability of Default (PD) models—where the goal is to separate riskier obligors or facilities from less risky ones—LGD models require validation to confirm their ability to discriminate between different levels of loss severity. This means verifying that the model assigns higher estimated LGD values to facilities that are likely to incur higher realised losses and lower estimated LGD values to those expected to incur lower losses.

One of the primary measures used to assess the discriminatory power of LGD models is the *generalised Area Under the Curve* (AUC). The generalised AUC is an extension of the classical AUC metric, which is typically applied in binary classification problems. In the context of LGD models, which deal with multi-class problems due to the continuous nature of loss severity, the generalised AUC provides a robust statistical tool for evaluation. It assesses the probability that, for any two randomly selected facilities, the one with the higher realised LGD was also assigned a higher estimated LGD by the model.

The process for conducting discriminatory power tests involves several steps:

- **Ordering Estimated LGD Values**: Facilities are ranked based on their estimated LGD values from the model, arranged from lowest to highest.
- **Discretising Realised LGD Values**: The actual realised LGD values are grouped into discrete categories corresponding to the estimated LGD grades or pools. These are then ordered similarly from low to high.
- **Comparing Rankings**: By comparing the rankings of estimated and realised LGD values, analysts can evaluate how effectively the model predicts the actual loss severity. A strong alignment between these rankings indicates good discriminatory power.

It's important to note that while this section focuses on LGD models, the principles and methods discussed are equally applicable to Credit Conversion Factor (CCF) models. Both types of models benefit from rigorous validation to ensure they accurately discriminate between different risk levels.

For a more detailed explanation of the statistical techniques and additional validation tools used in assessing discriminatory power, refer to Section 3.2 of the annex.

Generalized AUC (gAUC) The Generalized Area Under the Curve (gAUC) is a performance metric used to evaluate the discriminatory power of Loss Given Default (LGD) models in finance. It measures how effectively a model can differentiate between various levels of loss severity among obligors or facilities.

To calculate the gAUC, follow these steps:

1. **Segmentation of LGD**: Define facility grades or pools to segment the LGD values ordinally. These segments should be consistent with the institution's internal validation practices. If the LGD model utilizes more than 20 facility grades or pools, or if it is a continuous model, the data should be grouped into 12 predefined LGD segments based on specific criteria outlined by regulatory guidelines.

2. **Sample Preparation**: Gather information on the sample used for the initial gAUC calculation, including:

   - The time period of the validation sample (start date and end date).
   - The number of facilities included in the sample.
   - The variance observed within the validation sample.

3. **Calculation of Initial gAUC**: Compute the initial gAUC using the prepared validation sample. This initial value serves as a benchmark for future comparisons.

4. **Calculation of Current gAUC**: Calculate the gAUC for the current observation period using the same methodology as the initial calculation. This current gAUC reflects the model's present performance.

5. **Estimation of Standard Deviation**: Estimate the standard deviation of the current gAUC to assess the variability of the model's performance measure.

6. **Computation of Test Statistic**: Determine the test statistic by evaluating the difference between the current gAUC and the initial gAUC, adjusted by the estimated standard deviation. This statistic helps identify any significant changes in the model's discriminatory power.

7. **Determination of P-value**: Calculate the p-value as one minus the cumulative distribution function of the standard normal distribution evaluated at the test statistic. The p-value indicates the probability of observing such a result under the null hypothesis that there is no change in the model's performance.

8. **Analysis of Results**: Interpret the p-value to determine if the change in gAUC is statistically significant. A low p-value suggests that the model's performance has significantly changed, which may necessitate further investigation or model recalibration.

By systematically calculating and comparing the initial and current gAUC values, institutions can monitor the stability and effectiveness of their LGD models over time. This process is crucial for ensuring that the models remain accurate and reliable for risk assessment purposes.

For more detailed information on the calculation of the gAUC, please refer to Section 3.2 in the annex of the specification document.

7.7 Calibration Tests

Assessing the calibration of Loss Given Default (LGD) models is essential to ensure that the estimated LGD values accurately reflect the long-run average losses experienced by the institution. Proper calibration enhances the reliability of credit risk assessments and ensures compliance with regulatory standards.

Institutions should calibrate their LGD estimates to the long-run average LGD calculated in accordance with regulatory requirements. For this purpose, they should choose a calibration method that is appropriate for their LGD estimation methodology. Calibration should be conducted at the relevant level, such as each grade or pool, and additional calibration tests should be provided at the level of the relevant calibration segment.

When performing calibration tests, institutions should consider the following aspects:

- **Representativeness of Data**: Institutions should analyze the representativeness of the data used for calibration by considering not only the current characteristics of the portfolio but also any foreseeable changes due to specific actions or decisions already taken. Adjustments made on the basis of expected future changes should not lead to a decrease in the LGD estimates.
- **Treatment of Extremely High LGD Values**: If institutions observe extremely high values of realized LGDs, especially those significantly exceeding 100% for exposures with small outstanding amounts at the moment of default, they should identify relevant risk drivers to differentiate these observations. These specific characteristics should be adequately reflected in the assignment to grades or pools. When using a continuous rating scale in LGD estimation, creating a separate calibration segment for such exposures may be appropriate.
- **Selection of Calibration Methods**: Institutions should select calibration methods that align with their LGD estimation methodologies. Different approaches may be suitable depending on whether the LGD estimates are derived from statistical models, expert judgment, or a combination of both.
- **Use of Validation Tools**: To monitor the performance of LGD models in terms of predictive ability and calibration, institutions should employ appropriate validation tools. These tools help in assessing whether the LGD estimates are accurately predicting observed losses and facilitate timely adjustments to the models when necessary.

By thoroughly testing and validating the calibration of LGD models, institutions can ensure that their credit risk estimates remain robust and reflective of actual loss experiences. This process not only supports sound risk management practices but also demonstrates compliance with regulatory expectations regarding the estimation of LGD parameters.

7.7.1 LGD Backtesting (t-test)

The Loss Given Default (LGD) backtesting using the one-sample t-test for paired observations aims to assess the predictive ability of LGD estimates at the portfolio level, as well as at the grade, pool, or segment levels. This statistical test compares estimated LGDs with realised LGDs to determine if there is a significant difference between them.

Null Hypothesis (H_0): The estimated LGD is greater than or equal to the true (realised) LGD.

Alternative Hypothesis (H_1): The estimated LGD is less than the true (realised) LGD.

Assuming independent observations, the test statistic T is calculated as:

$$T = \frac{\bar{d}}{s_d/\sqrt{N}}$$

where:

- $\bar{d}$ is the mean difference between estimated and realised LGDs.
- s_d is the standard deviation of the differences.
- N is the number of facilities (observations).

Under the null hypothesis, T follows a Student's t-distribution with $(N-1)$ degrees of freedom.

The p-value is calculated as:

$$\text{p-value} = 1 - S_{N-1}(T)$$

where $S_{N-1}(T)$ is the cumulative distribution function (CDF) of the Student's t-distribution evaluated at T with $(N-1)$ degrees of freedom.

Steps to Perform LGD Backtesting (t-test):

1. **Calculate Differences:** For each facility, compute the difference between the estimated LGD and the realised LGD:

$$d_i = \text{Estimated LGD}_i - \text{Realised LGD}_i$$

2. **Compute Mean Difference:** Calculate the mean of the differences:

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$$

3. **Calculate Standard Deviation:** Compute the standard deviation of the differences:

$$s_d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2}$$

4. **Compute Test Statistic:** Determine the value of the test statistic:

$$T = \frac{\bar{d}}{s_d/\sqrt{N}}$$

5. **Determine p-value:** Evaluate the p-value using the t-distribution CDF:

$$\text{p-value} = 1 - S_{N-1}(T)$$

6. **Decision Rule:** Compare the p-value with the chosen significance level (e.g., $\alpha = 0.05$). If the p-value is less than α, reject the null hypothesis.

Interpretation:

- **Reject H_0:** If the p-value is less than the significance level, there is sufficient evidence to suggest that the estimated LGDs are significantly less than the realised LGDs.
- **Fail to Reject H_0:** If the p-value is greater than the significance level, there is not enough evidence to conclude that the estimated LGDs are less than the realised LGDs.

Python Implementation Example:

```
import numpy as np
from scipy import stats

# Estimated and realised LGD values (replace with actual data)
estimated_lgd = np.array([0.4, 0.5, 0.6, 0.3, 0.7])
realised_lgd = np.array([0.35, 0.55, 0.65, 0.25, 0.75])

# Calculate the differences between estimated and realised LGDs
d = estimated_lgd - realised_lgd

# Number of observations
N = len(d)

# Mean difference
d_bar = np.mean(d)

# Standard deviation of differences
s_d = np.std(d, ddof=1)

# Compute the test statistic
T = d_bar / (s_d / np.sqrt(N))

# Degrees of freedom
df = N - 1
```

```python
# Calculate the p-value (one-sided test)
p_value = 1 - stats.t.cdf(T, df=df)

# Output the results
print(f'Test statistic (T): {T:.4f}')
print(f'p-value: {p_value:.4f}')

# Determine if we reject the null hypothesis at alpha = 0.05
alpha = 0.05
if p_value < alpha:
    print('Reject the null hypothesis: Estimated LGD is significantly
        less than realised LGD.')
else:
    print('Fail to reject the null hypothesis: No significant
        difference between estimated and realised LGD.')
```

In this example:

- Replace the ‘estimated$_{l}gd`and`realised_{l}gd`arrayswiththeactualLGDestimatesandrealisations.-$
$Thecodecalculatesthetest statisticandp-value, andthendetermineswhethertorejectthenullhypothesisa$

Note: Ensure that the data used satisfies the assumptions of the t-test, namely that the differences are independently and identically distributed and approximately follow a normal distribution.

Calibration Plots Calibration plots are essential tools for assessing the accuracy and reliability of LGD estimates in credit risk modeling. They provide a visual comparison between the predicted LGD values and the actual observed losses, enabling institutions to evaluate the performance of their LGD models across different segments or grades.

Institutions should utilize calibration plots to:

- **Compare Predicted and Observed LGDs:** By plotting predicted LGD estimates against realized LGD values, institutions can identify patterns or discrepancies that may indicate areas where the model's predictive power could be improved.
- **Assess Long-Run Average LGD Alignment:** Calibration plots help ensure that LGD estimates are appropriately calibrated to the long-run average LGD, as calculated in accordance with Section 6.3.2. This alignment is crucial for maintaining consistency with historical loss experience.
- **Identify Extreme Loss Observations:** When extremely high realized LGD values occur—especially those exceeding 100% for exposures with small outstanding amounts at default—calibration plots can highlight these anomalies. Institutions should identify relevant risk drivers to differentiate these observations and consider creating separate calibration segments for such exposures.
- **Evaluate Data Representativeness:** Calibration plots aid in analyzing the representativeness of data by taking into account current portfolio characteristics and any foreseeable changes resulting from specific actions or decisions already taken. Adjustments based on expected future changes should be made cautiously and should not lead to decreased LGD estimates.

- **Provide Additional Calibration Tests:** Especially when LGD estimates are calibrated to the long-run average for each grade or pool, institutions should perform additional calibration tests at the level of relevant calibration segments to verify the robustness of their models.

By incorporating calibration plots into the validation process, institutions can enhance the transparency and effectiveness of their LGD models. This practice supports compliance with regulatory standards and contributes to a more accurate assessment of credit risk.

7.8 Coverage Level Adjusted R (CLAR)

The Coverage Level Adjusted R (CLAR) is a crucial concept in the calibration of Probability of Default (PD) estimates within financial institutions. It refers to the adjustments made to observed default rates to ensure that PD estimates accurately reflect the likely range of variability in default rates over different economic conditions.

Regulatory guidelines, such as those outlined in the Capital Requirements Regulation (CRR) and accompanying Guidelines (GL), emphasize the importance of representativeness in risk quantification. Specifically, adjustments should at least reflect the number of good and bad years in the available data relative to the period representative of the likely range of variability. This means that when bad years (periods with higher default rates) are underrepresented in the data, institutions should make upward adjustments to the observed average one-year default rates. Unless there is empirical evidence indicating that default rates are unrelated to economic conditions, neglecting this adjustment could lead to underestimation of risk.

The magnitude of the necessary adjustments depends on several factors:

- **Variability of Default Rates**: The greater the variability in default rates, the larger the adjustments required. A higher variability indicates more uncertainty and potential risk, warranting more conservative PD estimates.
- **Number of Good and Bad Years**: The adjustment should account for the proportion of good and bad years in the data. If bad years are underrepresented, the adjustment compensates for this imbalance.
- **Grade Assignment Dynamics**: In calibration at the grade or pool level, the dynamics of how exposures are assigned to grades can affect the adjustment. For instance, if grade assignments are sensitive to economic conditions, adjustments at the grade level become more significant.

The GL provides an overview of calibration methods applicable to different types of exposures (e.g., retail, non-retail, or purchased receivables) and clarifies that calibration can be performed at various levels:

- **Calibration Segment Level**: Adjustments are made considering the default rates at the segment level, which may encompass multiple grades or pools.

- **Grade or Pool Level**: Calibration is performed for each grade or pool individually, allowing for more granular PD estimates.
- **Portfolio Level**: When there is only one calibration segment, calibration at the segment level corresponds to the portfolio level.

To promote understanding and awareness of different calibration levels and to limit variability in Risk-Weighted Assets (RWA) stemming from these choices, the GL requires institutions to perform additional calibration tests at both the calibration segment level and the grade or pool level. For instance, if calibration is performed at the grade level, institutions should also test calibration at the segment level, and vice versa. This approach ensures that PD estimates are robust across different levels of aggregation.

It has been observed that, for certain exposure classes (e.g., retail exposures secured by immovable property), the adjustment from the observed average default rate to the final PD estimate is higher when calibration is performed at the grade or pool level compared to the portfolio level. This phenomenon occurs because grade-level calibration captures more detailed risk characteristics, necessitating larger adjustments to account for variability within each grade.

However, using a calibration sample that is not fully representative of the likely range of variability can result in PD estimates at the grade level that are lower than the Long-Run Average (LRA) default rate at that level. For example, if bad years are overrepresented in the calibration sample and the grade assignments are sensitive to economic conditions, the resulting adjustments might be biased, leading to underestimation of PDs in less risky grades.

In accordance with the CRR requirements on risk quantification, institutions must ensure that their PD estimates:

- Are based on data that is representative of the likely range of variability of default rates.
- Include appropriate adjustments to account for the representativeness of the data, especially regarding the occurrence of good and bad years.
- Reflect the inherent uncertainty and potential variability in default rates, incorporating a Margin of Conservatism (MoC) where necessary.

Understanding and applying the CLAR is essential for financial institutions to maintain compliance with regulatory standards and to ensure that their risk models accurately capture potential credit risk. By properly adjusting for coverage levels, institutions enhance the robustness of their PD estimates, leading to more reliable risk assessments and prudent capital allocation.

7.9 Stability Tests

Stability tests are essential for evaluating the consistency and reliability of internal ratings and risk parameters over a specific observation period. These tests help identify any significant changes in the performance of credit ratings and probability of default (PD) estimates, ensuring that the models remain robust over time.

Examples for analyzing the stability of internal ratings and risk parameters over a specific observation period for PD estimates can include:

- Comparing the distribution of rating grades at the beginning and end of the observation period.
- Performing statistical tests for each combination of rating grades from 1 to K to detect any significant shifts.
- Evaluating subsamples such as retail exposures, retail secured by immovable property non-SME, and retail exposures secured by qualifying revolving and other retail.

The test is performed for each combination of rating grades at the beginning and end of the relevant observation period. For instance, a statistical test can assess whether the migration between different rating grades is significant. This involves calculating test statistics and corresponding p-values to determine the likelihood of observed changes occurring by chance.

In a practical example, the test was performed for all subsamples listed in Table 19, including:

- Retail exposures
- Retail secured by immovable property non-SME
- Retail exposures secured by qualifying revolving and other retail

However, none of the tests showed statistical significance, indicating that the ratings remained stable over the observation period.

To facilitate the analysis of model design stability, it is important to compute test statistics and p-values. For cases where the rating grades are not equal, values for the test statistic and the p-value can be calculated. The p-value indicates the probability of observing a test statistic as extreme as, or more extreme than, the one observed under the null hypothesis.

- **Test Statistic**: Measures the extent of deviation between rating grades over time.
- **P-value**: Helps determine the statistical significance of the observed deviation.

Below is an example of how to perform a stability test using Python:

```
import pandas as pd
from scipy.stats import chi2_contingency

# Load the data: Assume 'ratings_start.csv' and 'ratings_end.csv'
    contain rating grades at the start and end
data_start = pd.read_csv('ratings_start.csv')
```

```
data_end = pd.read_csv('ratings_end.csv')

# Create frequency tables for the rating grades
freq_start = data_start['Rating'].value_counts().sort_index()
freq_end = data_end['Rating'].value_counts().sort_index()

# Combine the frequency tables into a contingency table
contingency_table = pd.DataFrame({
    'Start': freq_start,
    'End': freq_end
}).fillna(0)

# Perform the Chi-squared test
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print(f'Chi-squared statistic: {chi2:.2f}')
print(f'p-value: {p_value:.4f}')

# Interpret the result
alpha = 0.05
if p_value < alpha:
    print('The rating distributions are significantly different. Model
       instability detected.')
else:
    print('No significant difference in rating distributions. Model
       remains stable.')
```

By conducting these stability tests, financial institutions can ensure that their internal rating systems and risk parameters remain accurate and reliable, thereby supporting effective risk management and regulatory compliance.

Population Stability Index (PSI) The Population Stability Index (PSI) is a statistical metric used to measure the shift in the distribution of a variable between two different populations or over time. In the context of Loss Given Default (LGD) models, PSI is instrumental in monitoring the stability of LGD estimates and ensuring that the model remains reliable under changing economic conditions.

PSI quantifies the difference between the distribution of LGD predictions from the model's development sample and that of a new validation sample. A significant change in PSI indicates that the current portfolio's characteristics differ from those of the development population, which may impact the model's predictive power.

Monitoring PSI for LGD models is essential for several reasons:

- **Quantification of Downturn LGD:** The guidelines emphasize the need for appropriate quantification of downturn LGD. By tracking PSI, institutions can identify periods where the LGD distribution shifts, potentially signaling economic downturns. This enables institutions to adjust their LGD estimates to reflect adverse conditions accurately.
- **Regulatory Compliance and Harmonization:** Regulatory bodies aim to achieve convergence in the methodologies used for PD and LGD estimation to reduce unjustified variability in risk-weighted assets (RWA). Incorporating PSI analysis aligns

with these guidelines by providing a standardized approach to monitor model stability and performance over time.

- **Model Validation and Risk Management:** Regular PSI computation is a vital part of ongoing model validation processes. It helps detect model degradation early, allowing institutions to take corrective actions promptly. This proactive approach supports effective risk management and ensures that models remain robust under different economic scenarios.

Furthermore, understanding the correlation between default rates (PD) and LGD estimates is crucial, as the severity of losses during downturn periods may be higher. Some institutions define downturn periods based on the observed correlation between PD and LGD. By analyzing PSI alongside PD and LGD correlations, institutions can gain deeper insights into their risk profiles and adjust their models accordingly.

In line with the harmonization efforts and regulatory expectations, institutions should allocate sufficient time and resources to implement PSI monitoring frameworks. This includes adjusting for the phasing-in approach and meeting implementation deadlines set by regulatory authorities. By doing so, institutions not only comply with guidelines but also enhance the reliability of their risk assessments in a changing economic environment.

7.10 Qualitative Validation

The qualitative validation of Loss Given Default (LGD) models is a critical component of the overall model validation framework. It focuses on assessing the appropriateness of the methodologies, assumptions, and processes used in estimating LGD, ensuring that they are suitable for the portfolio under consideration and comply with regulatory requirements.

Qualitative validation aims to ensure that LGD estimates are correctly assigned to the application portfolio and that the distribution of LGD across different facilities accurately reflects their inherent risk characteristics. This involves a comprehensive analysis of the model design, data quality, estimation techniques, and the implementation of model outputs.

Key aspects of qualitative validation include:

- **Assessment of Methodological Choices:** The validation function is expected to critically evaluate the methodological choices made in deriving LGD best estimates. This includes assessing the appropriateness of the estimation methods in relation to the long-run average realised LGD per grades or pools, and challenging the assumptions and techniques used.
- **Review of Data Quality and Completeness:** Ensuring the accuracy and completeness of the data used in the LGD model is essential. The validation should assess the quality of historical recovery data, the handling of incomplete cases, and the estimation of future recoveries on open defaults.
- **Appropriateness of LGD Assignments:** The validation should verify that LGD estimates are appropriately assigned to facilities within the portfolio. This involves

checking that the risk characteristics of exposures are accurately captured and that LGD values are reflective of these risks.

- **Analysis of LGD Distribution:** Evaluating the distribution of LGD across different grades or pools helps identify any inconsistencies or anomalies. The validation function should ensure that the LGD model differentiates effectively between varying levels of risk.

- **Review of Back-Testing Processes:** The validation is expected to examine the back-testing procedures for LGD estimates. This includes performing checks that compare LGD estimates with realised LGDs using only closed cases, as well as incorporating estimations of future recoveries on incomplete cases. Additionally, comparing the estimations of future costs and recoveries with their actual realisations is crucial.

- **Assessment of Downturn LGD Estimation:** For downturn LGD estimates, the validation should assess the methodology chosen for estimation, compare it with long-run averages, and evaluate the sensitivity of the estimates to changes in economic cycles. This includes reviewing the aggregation of impacts from intermediate parameters and the processes used for defaulted exposures.

- **Compliance with Regulatory Standards:** The validation function must ensure that the LGD model complies with all relevant regulatory requirements. This involves verifying adherence to guidelines on estimation methods, data use, and validation practices prescribed by regulatory bodies.

- **Identification and Analysis of Deficiencies:** In cases where the realised LGD in a grade or pool falls outside the expected range, the validation function is expected to analyse the deficiency thoroughly. This includes investigating possible reasons for the discrepancy and determining whether it indicates a model weakness that needs to be addressed.

- **Documentation and Transparency:** Comprehensive documentation of the LGD model and validation processes is essential. The validation function should ensure that all methodological choices, assumptions, and validation findings are well-documented and transparent to stakeholders.

By focusing on these qualitative aspects, the validation function can provide a robust challenge to the LGD model, ensuring that it is not only statistically sound but also conceptually appropriate. This holistic approach strengthens the reliability of LGD estimates, supports effective risk management, and enhances compliance with regulatory expectations.

Assignment Process Statistics Assignment process statistics are essential for the qualitative validation of Loss Given Default (LGD) models. They provide valuable insights into how LGD estimates are assigned across the application portfolio, highlighting potential inconsistencies or issues in the estimation process. A critical aspect of this analysis is examining the relative frequency of LGD estimates, particularly focusing on cases with missing values or those forced to default values.

By analyzing the distribution of LGD assignments, validation teams can ensure that the LGD model accurately reflects the risk characteristics of the portfolio. High frequencies of missing or overridden LGD estimates may indicate deficiencies in the model's ability to capture certain risk factors or issues with data quality. Identifying and addressing these issues is crucial for maintaining the integrity of the LGD estimates and ensuring regulatory compliance.

In addition to qualitative assessments, assignment process statistics should be complemented with quantitative tools such as back-testing of LGD estimates. Back-testing involves comparing the assigned LGD estimates with the realized LGDs to evaluate the predictive accuracy of the model. The validation process is expected to perform the following checks:

1. **Closed Cases Comparison:** Comparing LGD estimates with realized LGDs using only closed cases, including those where the maximum period of the recovery process has been reached.
2. **All Cases Comparison:** Incorporating all cases by estimating future recoveries on incomplete cases and comparing these estimates with the assigned LGD values.
3. **Estimation Accuracy Analysis:** Analyzing the accuracy of estimations for future costs and recoveries on incomplete cases against their actual realizations.

If the realized LGD in a specific grade or pool falls outside the expected range, the validation function should analyze the deficiency. This involves assessing the methodological choices made in deriving the LGD best estimates, particularly in relation to the long-run average realized LGDs per grade or pool. Challenging these methodological choices ensures that the LGD model remains robust and continues to provide accurate risk assessments.

Ultimately, the analysis of assignment process statistics ensures that LGD estimates are appropriately assigned across the application portfolio. It aids in identifying areas where the model may require refinement and supports the continuous improvement of risk management practices within the institution.

Portfolio Distribution The analysis of the portfolio distribution is a crucial component in the qualitative validation of *Loss Given Default* (LGD) models. It ensures that LGD estimates are appropriately assigned to different segments of the application portfolio and that the distribution of LGD values across these segments is reasonable and reflective of actual observed losses.

An effective validation process involves examining the LGD application portfolio distribution by facility grade or pool, or by predefined LGD segments. This analysis helps identify any inconsistencies or anomalies in the allocation of LGD estimates across various risk categories. By scrutinizing the distribution, validators can assess whether the LGD model adequately captures the variations in loss severity associated with different borrower characteristics, collateral types, or facility features.

Comparing LGD estimates with realised LGDs serves as an additional check on the model's performance. This comparison can reveal discrepancies between expected and actual losses, indicating potential deficiencies in the model's design or calibration. Such discrepancies may suggest that the model does not fully account for factors influencing recovery rates or that it may not be responsive to changes in economic conditions.

The qualitative validation aims to ensure that the LGD estimates are not only statistically robust but also economically meaningful and operationally sound. By thoroughly analyzing the portfolio distribution, the validation function can verify that the LGD model aligns with regulatory expectations and supports sound risk management practices.

8 EAD and CCF Model Validation

The validation of Exposure at Default (EAD) and Credit Conversion Factor (CCF) models is a critical component of regulatory compliance in finance. EAD represents the estimated amount that a borrower will owe at the time of default, while CCF is a factor used to convert off-balance sheet exposures into credit exposure equivalents.

In validating these models, it is essential to consider *facilities with missing EAD or CCF estimates*. These are facilities that do not have a CCF or EAD estimate at a given point in time but fall within the scope of the model under consideration. It is important to note that this term does not include facilities whose CCF or EAD estimates are based on missing or partly missing information.

Facilities covered by an EAD approach are those that are underpinned by a direct EAD estimate, such as estimates in the "region of instability." Proper identification and inclusion of these facilities are vital for an accurate assessment of credit risk.

The portfolio's exposure at default (EAD) is defined as the estimated exposure after applying the CCF, in accordance with Article 166 of the Capital Requirements Regulation (CRR), which pertains to Probability of Default (PD), Loss Given Default (LGD), CCF models, and the slotting approach.

An important aspect of EAD validation is the aggregation of exposures. This involves summing the estimated exposure at default and the observed drawings. Additionally, exposure should be weighted by the committed but undrawn credit amount to accurately reflect the potential risk.

In conclusion, rigorous validation of EAD and CCF models ensures that credit risk is appropriately measured and managed, thereby contributing to the overall stability and compliance of financial institutions.

8.1 Introduction to EAD/CCF

Exposure at Default (**EAD**) and Credit Conversion Factor (**CCF**) are fundamental concepts in credit risk management and regulatory compliance within the financial industry. EAD represents the anticipated amount of exposure to a counterparty at the time of default. It is a crucial parameter for calculating the capital requirements under regulatory frameworks such as the Capital Requirements Regulation (**CRR**). According to Article 166 of the CRR, EAD is defined as the estimated exposure after applying the CCF to off-balance sheet items.

CCF is a risk parameter used to convert off-balance sheet exposures, such as undrawn credit lines and letters of credit, into credit exposure equivalents. It reflects the likelihood that an off-balance sheet item will convert into an on-balance sheet exposure before or at the time of default. The accurate estimation of CCF is essential for determining EAD, as it directly impacts the assessment of potential losses.

Validating EAD and CCF models presents specific challenges for financial institutions:

- **Facilities with Missing Estimates:** One challenge involves facilities that do not have a CCF or EAD estimate at a given point in time but are within the scope of the model. These cases require careful consideration to ensure that the absence of estimates does not lead to underestimation of risk. Notably, this issue excludes facilities where estimates are based on missing or partially missing information.

- **Predictive Ability Analysis:** Assessing the predictive ability, or calibration, of the CCF parameter is crucial. The goal is to ensure that the CCF enables an accurate prediction of EAD. For facilities covered by a specific EAD approach, a simplified analysis may be applied, but it must still provide assurance of the model's predictive performance.

- **Alignment with LGD Calculations:** There is a need for consistency between the EAD considered for CCF purposes and the EAD used in the denominator of realised Loss Given Default (**LGD**). Inconsistent approaches can lead to inaccurate risk assessments. Studies have indicated that while some retail models maintain adequate alignment, others fail to do so, particularly regarding additional (after default) drawings. This misalignment can result in an inconsistent treatment of exposures and potential underestimation of losses.

- **Integration of Machine Learning Models:** The increasing use of Machine Learning (**ML**) models in estimating risk parameters—including Probability of Default (**PD**), LGD, Expected Loss Best Estimate (**ELBE**), EAD, and CCF—introduces additional validation complexities. ML models may offer enhanced predictive capabilities but also require robust validation frameworks to address issues such as explainability, overfitting, and compliance with regulatory expectations.

Addressing these challenges is vital for the integrity of credit risk models. Effective validation ensures that EAD and CCF estimates are reliable and that the institution remains compliant with regulatory standards. It also enhances the risk management framework by providing more accurate assessments of potential exposures, thereby supporting better decision-making and capital allocation.

8.2 CCF Discriminatory Power Tests

The assessment of discriminatory power aims to ensure that Credit Conversion Factor (CCF) models are capable of distinguishing between facilities with high and low CCF values. This is crucial for accurate risk differentiation and effective credit risk management.

In this context, we utilize the *generalized Area Under the Curve (AUC)* as the measure to assess the discriminatory power of CCF models. The generalized AUC is an extension of the classical AUC, adapted to handle multi-class classification problems commonly encountered in CCF modeling. This measure evaluates the ability of the model to correctly rank facilities according to their CCF values.

Analyses of discriminatory power for CCF models are designed to ensure that the model effectively discriminates between facilities with high CCF values and those with low CCF values. This approach mirrors the assessment performed for Probability of Default (PD) models, where the ranking of obligors or facilities is evaluated to appropriately separate riskier from less risky ones.[7]

Somers' D is another statistical measure employed in this context, calculated based on the methodologies proposed by Brown and Benedetti (1977)[8] and further elaborated by Göktaş and İşçi (2011).[9] These methodologies provide a robust framework for evaluating the strength and direction of association between predicted and observed CCF values.

More detailed information on the statistical methods and computations referred to in this section can be found in the annex, Section 3.2. Although the annex primarily discusses Loss Given Default (LGD) models, the results and methodologies are equally applicable to CCF models.

By effectively assessing the discriminatory power of CCF models using these statistical tools, financial institutions can enhance their credit risk models, leading to more accurate capital requirements and better risk management practices.

Generalized AUC (gAUC)

The assessment of discriminatory power aims to ensure that Credit Conversion Factor (CCF) models are capable of distinguishing between facilities with high and low CCF values. To evaluate this discriminatory power, we use the generalized Area Under the Curve (gAUC), a validation tool based on a generalization of the classical AUC that can be applied to multi-class problems.

The gAUC is calculated based on facility grades or pools, which serve as the ordinal segmentation of Loss Given Default (LGD). These facility grades or pools are defined consistently with the institution's internal validation process. If the model is based on more than 20 facility grades or pools, or if it is a continuous LGD model, the test is performed using 12 predefined "LGD segments" according to specific criteria.

At the time of the initial validation, the gAUC is referred to as $gAUC_{init}$, and for the relevant observation period, it is referred to as $gAUC_{curr}$. The estimated standard deviation of $gAUC_{curr}$ is denoted as s, which is the square root of the variance.

For more detailed information on the calculation of the gAUC and the statistics mentioned above, please refer to the specification in the annex, Section 3.2.

8.3 CCF Calibration Tests

The aim of the validation tools outlined in this section is to monitor CCF models' performance in the following areas of investigation:

[7] Analyses of discriminatory power for PD models should be designed to ensure that the ranking of obligors/facilities resulting from the rating methodology appropriately separates riskier and less risky obligors/facilities. Similarly, analyses of discriminatory power for LGD (respective CCF) models should be designed to ensure that the LGD (respectively CCF) model is able to discriminate between facilities with high values and those with low values.

[8] Brown, M. and Benedetti, J. (1977). *Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables.* Journal of the American Statistical Association, Vol. 72, pp. 309–315.

[9] Göktaş, A., and İşçi, Ö. (2011). *A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation.* Metodološki zvezki, 8(1), pp. 17–37.

- **Predictive ability (calibration) of CCF models.**
- **Qualitative aspects of CCF models.**

Predictive Ability of CCF Models

The analysis of predictive ability, or calibration, aims to ensure that the Credit Conversion Factor (CCF) risk parameter facilitates a good prediction of Exposure at Default (EAD). This involves assessing whether the estimated CCF values accurately predict the actual utilization of credit lines at the time of default.

An illustration of the back-testing sample used for CCF models is shown in Figure **??**. This figure demonstrates the construction of the sample used to assess the predictive ability of CCF models. It uses the example of two generic facilities (A and B) whose recovery process has begun in the relevant one-year observation period, indicating the estimated CCF that is to be back-tested in each case.

[Illustration of the back-testing sample used for CCF models]

Where facilities are covered by an EAD approach (see point (<>)(g) of Section (<>)2.9.1), a simplified analysis is applied.

Qualitative Aspects of CCF Models

The analysis of qualitative aspects of CCF models is aimed at ensuring the appropriateness of the assignment process for this parameter. This involves reviewing the methodologies, assumptions, and data used in the CCF model to ensure they are appropriate and comply with regulatory requirements.

8.3.1 CCF Backtesting (t-test)

The objective of the CCF backtesting using a one-sample t-test for paired observations is to assess the predictive ability of Credit Conversion Factor (CCF) estimates at the facility grade or pool level, in accordance with Article 182(1)(a) of the Capital Requirements Regulation (CRR).

This backtesting tool compares the estimated CCFs with the realised CCFs under the null hypothesis that the estimated CCF is greater than or equal to the true CCF. This is a one-sided hypothesis test, assuming independent observations.

The one-sample t-test for paired observations follows a Student's t-distribution with degrees of freedom equal to the number of facilities in the backtesting sample, excluding those facilities covered by an Exposure at Default (EAD) approach, minus one.

When constructing the backtesting sample, facilities are included if defaults occur in the relevant one-year observation period. The estimated CCF to be backtested for an individual facility relates to either:

- the point in time one year before the facility's default; or
- the beginning of the relevant one-year observation period.

Figure 8 illustrates the construction of the CCF backtesting sample for two generic facilities.

8.4 EAD Model Validation

Validating Exposure at Default (EAD) models is a critical component in credit risk management, particularly when direct EAD estimates are employed. Facilities that are covered by direct EAD estimates often reside in the *region of instability*, where exposure amounts can be highly volatile and uncertain. The validation of these models ensures that the estimates accurately reflect potential exposures in the event of default.

Assessment of Model Changes Due to Policy Implementation

Recent surveys have enabled a direct assessment of how new regulatory policies impact EAD models within the Internal Ratings-Based (IRB) approach. Specifically, for selected questions, it was possible to determine:

- The proportion of models that would be affected by the chosen policy.
- The share of exposure amounts requiring changes in modelling practices once new guidelines (GLs) come into force.

Table 32 from the survey provides an overview of the resulting model changes, highlighting both the share of Probability of Default (PD) models impacted (%) and the share of exposure values covered by these PD models (% EAD). This analysis is crucial for institutions to understand the extent of adjustments needed to comply with upcoming regulatory requirements.

Adjustments to Observed Default Rates

Institutions often apply adjustments to observed average default rates (DR) when estimating PD to enhance model accuracy. Figure 5 from the survey indicates:

- Adjustments are applied in approximately half of the models and exposures assessed.
- A higher proportion of adjustments occurs in retail models compared to corporate models.

The validation process must scrutinize these adjustments to ensure they are justified and improve the predictive power of the models without introducing bias.

Validation Steps for Direct EAD Estimates

When validating EAD models with direct estimates, the following steps are essential:

1. *Data Integrity Verification*: Ensure that the data used for estimation is accurate, complete, and representative of the current portfolio.
2. *Model Assumption Review*: Examine the assumptions underlying the EAD model to confirm their validity in the context of the facilities being assessed.

3. *Back-Testing Analysis*: Compare predicted EAD values against actual exposures at default to evaluate model performance.
4. *Impact Assessment of Policy Changes*: Analyze how new GLs affect model outputs and determine if recalibration is necessary.
5. *Comparison Across Models*: Benchmark the EAD estimates against other models or industry standards to identify inconsistencies or areas for improvement.

Consideration of Retail vs. Corporate Models

The validation process should account for differences between retail and corporate models. The survey findings suggest:

- Adjustments are more prevalent in retail models due to the higher variability and granularity of retail exposures.
- Corporate models may require different validation techniques owing to the unique characteristics of corporate exposures.

Understanding these distinctions is vital for tailoring the validation approach appropriately.

Conclusion

Effective validation of EAD models with direct estimates is indispensable for managing credit risk and ensuring regulatory compliance. By thoroughly assessing model changes necessitated by new policies, scrutinizing adjustments to default rates, and considering the specific attributes of different exposure types, institutions can enhance the reliability of their EAD estimates. This rigorous validation process ultimately supports better risk management decisions and strengthens the resilience of financial institutions in the face of default events.

EAD Backtesting (t-test) EAD backtesting using the t-test is a statistical method employed to evaluate the accuracy of Exposure at Default (EAD) estimates by comparing them with the actual drawn amounts at the time of default. The primary objective is to determine whether there is a significant difference between the predicted EAD values and the realised exposures when defaults occur.

The t-test is conducted under the null hypothesis that the estimated EAD values are equal to the realised drawn amounts at default. This implies that any observed differences are due to random variation rather than systematic bias in the EAD estimation model. The test follows these steps:

1. **Calculate Differences:** For each defaulted facility, compute the difference between the realised drawn amount at default and the corresponding estimated EAD.
2. **Compute Mean Difference:** Determine the average of these differences across all defaulted facilities. This mean difference reflects the overall bias in the EAD estimates.

3. **Estimate Variance:** Calculate the variance of the differences to assess the dispersion around the mean difference. This variance is used to estimate the standard error of the mean difference.

4. **Calculate Test Statistic:** Compute the t-test statistic by dividing the mean difference by its standard error. This statistic measures how many standard errors the mean difference is away from zero.

5. **Determine p-value:** Obtain the p-value associated with the test statistic using the cumulative distribution function of the Student's t-distribution. The degrees of freedom are equal to the number of defaulted facilities minus one.

A low p-value indicates that the null hypothesis can be rejected, suggesting there is a statistically significant difference between the estimated EADs and the realised drawn amounts. This outcome implies that the EAD estimation model may not be accurately capturing the exposure at default and may require recalibration or adjustment.

It is important to note that this t-test assumes independent observations and that the differences between estimated and realised values are approximately normally distributed. The test provides a quantitative framework for assessing the performance of the EAD estimation process and identifying potential areas for improvement in the risk management models.

8.5 CCF/EAD Stability Test

The Credit Conversion Factor (CCF) and Exposure at Default (EAD) are crucial parameters in credit risk modeling, particularly for portfolios with contingent exposures. Ensuring the stability of these parameters over time is essential for accurate risk assessment and regulatory compliance. This subsection explains the methods used to test the stability of CCF/EAD estimates and demonstrates how to perform these tests.

Purpose of the Stability Test

The primary goal of the CCF/EAD stability test is to analyze whether internal estimates remain consistent over a specific observation period. Stability indicates that the risk parameters are reliable and that the credit risk models are robust against changes in economic conditions.

Approach to Testing Stability

To assess the stability of CCF/EAD estimates, institutions can perform the following analyses:

- *Analysis of Model Design Stability*: Evaluate the consistency of the model's structure, variables, and assumptions over time.

- *Comparison of Estimates Over Time*: Compare CCF/EAD estimates at the beginning and end of the observation period for each rating grade.

- *Statistical Testing*: Use statistical tests to determine if observed differences in estimates are significant.

Performing the Stability Test

The test is performed for each combination of rating grades 1 to K at the beginning and end of the relevant observation period. The steps involved are:

1. *Data Collection*: Gather CCF/EAD data segmented by rating grades for two points in time.
2. *Statistical Analysis*: Calculate the differences in estimates and perform statistical tests to assess significance.
3. *Interpretation*: Analyze the p-values from the tests to determine stability.

Example: Stability Analysis Using Python

Below is an example of how to perform a stability test for CCF estimates across different rating grades using Python.

```
import pandas as pd
from scipy.stats import norm

# Load CCF data for the beginning and end of the observation period
ccf_start = pd.read_csv('ccf_start_period.csv')  # Data at the
    beginning
ccf_end = pd.read_csv('ccf_end_period.csv')      # Data at the end

# Define rating grades
rating_grades = ccf_start['RatingGrade'].unique()

# Perform stability test for each rating grade
for grade in rating_grades:
    # Extract CCF estimates for the current rating grade
    ccf_start_grade = ccf_start[ccf_start['RatingGrade'] == grade]['CCF
        ']
    ccf_end_grade = ccf_end[ccf_end['RatingGrade'] == grade]['CCF']

    # Calculate mean and standard deviation
    mean_start = ccf_start_grade.mean()
    std_start = ccf_start_grade.std()
    mean_end = ccf_end_grade.mean()
    std_end = ccf_end_grade.std()

    # Calculate z-statistic for the difference in means
    n_start = len(ccf_start_grade)
    n_end = len(ccf_end_grade)
    pooled_std = ((std_start**2)/n_start + (std_end**2)/n_end)**0.5
    z_stat = (mean_start - mean_end) / pooled_std

    # Calculate p-value
    p_value = 2 * (1 - norm.cdf(abs(z_stat)))

    # Output the results
    print(f"Rating Grade: {grade}")
    print(f"Mean CCF at Start: {mean_start:.4f}")
    print(f"Mean CCF at End: {mean_end:.4f}")
    print(f"Z-statistic: {z_stat:.4f}")
    print(f"P-value: {p_value:.4f}")
```

```
if p_value < 0.05:
    print("Result: Significant difference detected.\n")
else:
    print("Result: No significant difference detected.\n")
```

Interpretation of the Results

In the above code:

- We calculate the mean CCF for each rating grade at the beginning and end of the observation period.
- The z-statistic measures the difference between the two means relative to the variability observed.
- The p-value indicates the probability of observing such a difference if there were no real change (i.e., if the estimates are stable).
- A p-value less than 0.05 suggests a statistically significant difference, indicating potential instability in the CCF estimates for that rating grade.

Conclusion

Conducting stability analyses of CCF/EAD estimates over time is vital for validating the consistency and reliability of credit risk models. By performing statistical tests for each rating grade, institutions can identify significant changes in risk parameters and take appropriate actions to adjust their models or reconsider their assumptions.

Population Stability Index (PSI) The Population Stability Index (PSI) is a statistical measure used to assess changes in the distribution of a portfolio over time. For facilities covered by a Credit Conversion Factor (CCF) approach, the PSI is calculated based on the number of facilities in the application portfolio at the beginning and end of the observation period. This measure helps in identifying shifts in the portfolio that may impact the predictive accuracy of credit risk models.

Monitoring the PSI is crucial to ensure that the CCF risk parameter continues to provide a reliable prediction of Exposure at Default (EAD). Significant changes in the PSI may indicate that the underlying assumptions of the model are no longer valid, and adjustments may be necessary.

In cases where facilities are covered by an EAD approach, a simplified analysis is applied. This is because the primary goal is to ensure that the CCF risk parameter facilitates an accurate prediction of EAD without the need for complex adjustments.

When calculating the PSI, it is important to exclude facilities covered by the EAD approach from the analysis. The focus should be on the facilities relevant to the CCF approach to maintain the accuracy and relevance of the PSI.

Understanding and applying the PSI allows financial institutions to detect early signs of changes in their credit portfolios. This proactive monitoring supports better risk management and compliance with regulatory requirements.

8.6 Qualitative Validation

The qualitative validation of Credit Conversion Factor (CCF) and Exposure at Default (EAD) models focuses on ensuring that these models are applied within their approved scope, in accordance with Article 143(3) of the Capital Requirements Regulation (CRR). This involves comparing facility types and characteristics to confirm that they align with the intended range of application for the CCF and EAD models.

An essential aspect of this validation is the analysis of the application portfolio. This analysis aims to assess the distribution of estimated CCFs and monitor their evolution over time. By evaluating the distribution, we can identify trends or deviations that may indicate model performance issues or shifts in portfolio composition. For facilities covered by an EAD approach, a simplified analysis is conducted due to the different nature of EAD estimation compared to CCF estimation.

Another critical component is the assessment of facilities with missing CCF or EAD estimates. These are facilities that, at a given point in time, do not have an assigned CCF or EAD estimate but fall within the scope of the model under consideration. It is important to note that this does not include facilities whose estimates are based on incomplete or partially missing information. Statistics on the frequency of such facilities should be compiled to evaluate the effectiveness of the CCF assignment process and identify any potential gaps in model coverage.

Furthermore, the analysis of predictive ability, also known as calibration, is performed to ensure that the CCF risk parameter provides accurate predictions of EAD. For facilities covered by an EAD approach, a simplified calibration analysis is applied. This step verifies that the model's estimates are reliable and that the model remains robust over time.

In summary, qualitative validation encompasses:

- Ensuring the model's range of application aligns with the approved scope by comparing facility types and characteristics.
- Assessing the distribution and evolution over time of estimated CCFs within the application portfolio.
- Compiling statistics on facilities with missing CCF or EAD estimates to evaluate the CCF assignment process.
- Analyzing the predictive ability of the CCF risk parameter to confirm accurate EAD predictions, with a simplified approach for facilities under the EAD model.

This comprehensive approach helps maintain the integrity and reliability of CCF and EAD models, ensuring they continue to meet regulatory requirements and provide accurate risk assessments.

Assignment Process Statistics Assignment process statistics are essential for the qualitative validation of rating systems in finance. They provide valuable insights into the integrity and effectiveness of the rating assignment process, especially when qualitative assessments, expert-based estimates, and assumptions play a significant role.

The validation process should assess the performance of the rating systems by employing both qualitative and quantitative methods. This assessment focuses on two main aspects:

- **Ranking Power**: Evaluating how well the rating system ranks borrowers according to their creditworthiness.
- **Calibration Appropriateness**: Assessing the accuracy of risk parameter estimations used in rating assignments.

An illustration of the data basis for qualitative rating process statistics is crucial. It helps in understanding the composition of the back-testing sample and the underlying data used for validation. This illustration often includes:

- The sources of data collected for the rating process.
- The methodology used for processing and analyzing the data.
- The criteria for selecting the back-testing sample.

Implementation of assignment process statistics involves calculating summary statistics at the portfolio level for various process deficiencies. These deficiencies may include:

- Instances of non-compliance with established rating guidelines.
- Occurrences of overrides in the rating assignment process without proper justification.
- Delays in updating ratings following significant changes in borrower circumstances.
- Inconsistencies in the application of expert judgment across different cases.

By identifying and analyzing these deficiencies, institutions can enhance the integrity of their rating assignment processes. Regular monitoring and reporting of assignment process statistics ensure that the rating systems remain robust, reliable, and aligned with regulatory compliance requirements.

8.6.1 Portfolio Distribution

The qualitative validation of a portfolio model necessitates a comprehensive understanding of the portfolio's distribution at the time of risk parameter estimation. This begins with analyzing the *application portfolio*, which refers to the actual portfolio of exposures within the range of application of the Probability of Default (PD) or Loss Given Default (LGD) model at the time of estimation.

All summary statistics in this section are computed based on the portfolio's composition at the beginning of the observation period. This approach ensures that the analysis accurately reflects the portfolio's characteristics without the influence of subsequent data exclusions or adjustments. Specifically, the size of the portfolio (M) is considered before any data exclusions to capture the complete risk profile.

Key aspects of the portfolio distribution include:

- **Portfolio Size (M):** Total number of exposures in the application portfolio prior to data exclusions. A larger M provides a more robust statistical foundation for validation purposes.
- **Summary Statistics Computation:** Calculations are based on the portfolio's composition at the reference point for estimation (see Section 2.9.1). This includes measures such as mean, median, variance, and other relevant statistical indicators.
- **Distribution of Estimated CCF:** The analysis focuses on assessing the distribution of the estimated Credit Conversion Factors (CCF) and its evolution over time. Monitoring changes in the CCF distribution helps in understanding shifts in exposure at default.
- **Simplified Analysis for EAD Approaches:** For facilities covered by an Exposure at Default (EAD) approach within the application portfolio, a simplified analysis is performed. This ensures that the validation process remains efficient while still capturing essential risk characteristics.

By thoroughly examining these qualitative aspects, the validation process aims to ensure that the risk parameters estimated by the model are appropriate for the portfolio's actual composition. This approach enhances the reliability of the model's predictions and supports compliance with regulatory requirements.

9 ELBE and LGD-in-default Validation

The validation of Expected Loss Best Estimate (ELBE) and Loss Given Default (LGD) in-default models is a critical aspect of credit risk management under the Internal Ratings-Based (IRB) approach. The Guidelines (GLs) clarify that, unless otherwise specified, all provisions applicable to LGD estimation for non-defaulted exposures also apply to LGD in-default and ELBE estimates. This alignment is intended to minimize cliff effects and ensure consistency in risk measurement across defaulted and non-defaulted exposures.

One significant area addressed by the GLs is the estimation approaches permitted for ELBE and LGD in-default. Institutions are required to estimate ELBE using an LGD model similar to that used for non-defaulted exposures, calibrated to current economic conditions and incorporating all relevant post-default information. Consequently, the practice of using accounting provisions as proxies for ELBE estimates—which currently accounts for 26% of models—is no longer permissible unless these provisions stem from models that comply with the specified conditions. It is anticipated that around 63% of ELBE models will need to be updated to meet these requirements.

The GLs also emphasize the importance of grouping defaulted exposures based on observed recovery patterns. This necessitates careful selection of reference dates to accurately capture the timing and magnitude of recoveries. By analyzing recovery patterns, institutions can enhance the accuracy of their LGD in-default and ELBE estimates, ultimately leading to more reliable assessments of expected and unexpected losses.

When overrides to model outputs are warranted, institutions must ensure consistency between ELBE and LGD in-default estimates. Specifically, any reasons for overriding ELBE outputs that are also relevant to LGD in-default should be consistently applied. This ensures that the add-on to the ELBE adequately covers potential increases in loss rates due to additional unexpected losses during the recovery period, in accordance with Article 181(1)(h) of Regulation (EU) No 575/2013.

It is important to note that while ELBE and LGD in-default models are closely related, they may not necessarily have the same scope. During back-testing, estimated LGD in-default should be analyzed at the ELBE portfolio level, as well as at the level of grades, pools, or segments. This detailed analysis aids in identifying any discrepancies and ensures that the models accurately reflect the risk characteristics of different segments within the defaulted exposure portfolio.

In conclusion, the validation of ELBE and LGD in-default models demands adherence to the updated GLs, which call for rigorous modeling practices and consistent application of policies. Institutions must reassess their current models and estimation approaches to ensure compliance, enhance the reliability of their risk estimates, and maintain robust credit risk management frameworks.

9.1 Introduction

Effective credit risk modeling is essential for financial institutions to assess and manage the potential losses associated with borrower defaults. Two critical components in this domain are the *Expected Loss Best Estimate* (ELBE) and the *Loss Given Default* for defaulted assets, commonly referred to as LGD-in-default. These measures play a vital role in quantifying the expected and unexpected losses on defaulted exposures, thereby influencing capital requirements and provisioning.

ELBE represents the institution's best estimate of the expected loss on a defaulted exposure, taking into account all available information about the borrower and the specific characteristics of the exposure at the time of default. It is a forward-looking measure that reflects anticipated recovery rates and the timing of recoveries. Conversely, LGD-in-default measures the potential loss severity on a defaulted exposure, incorporating an additional unexpected loss component to account for uncertainties beyond the expected loss.

Recent investigations into modeling practices have revealed diverse approaches among institutions in estimating ELBE and LGD-in-default. Key observations include:

- In 29% of cases, institutions had a dedicated ELBE model in place.
  - Of these, 62% based their expected loss estimation on the LGD performing model.
  - The majority utilized empirical evidence grounded in internal data for ELBE estimation.
- In 44% of cases, ELBE was set equal to the specific credit risk adjustments for the exposure.
- In the remaining cases, institutions lacked dedicated ELBE models, often defaulting to other estimation methods.

Concerning LGD-in-default:

- Both ELBE and LGD-in-default were assigned to defaulted exposures in 60% of cases.
- Only ELBE was calculated in 13% of cases.
- Only LGD-in-default was calculated in 5% of cases.
- Neither ELBE nor LGD-in-default values existed in the remaining 22% of cases.

These findings highlight several issues:

- **Lack of Dedicated Models:** Approximately 20% of cases revealed the absence of dedicated ELBE or LGD-in-default models.
- **Weaknesses in Modeling Approaches:** Around one-third of cases exhibited shortcomings in the modeling methodologies employed.
- **Insufficient Justification and Documentation:** Institutions often lacked justification for the assumptions underpinning their estimations and did not provide clear documentation on the breakdown of ELBE and LGD-in-default or the unexpected loss add-on component when used.

It is also noteworthy that ELBE calculations are typically performed using the ELBE model in place at the end of the observation period but based on customer and facility-specific input data (such as collateral valuations and risk factors) relevant to the estimation date. This approach ensures that ELBE reflects current information pertinent to the exposure under consideration. However, discrepancies may arise when the scopes of ELBE and LGD-in-default models differ, leading to individual ELBE values originating from different models.

The variability in practices underscores the importance of establishing robust, well-justified, and transparent modeling frameworks for ELBE and LGD-in-default. Such frameworks should be grounded in empirical evidence and leverage internal data to enhance reliability. Furthermore, clear documentation and justification of modeling assumptions are essential for regulatory compliance and for stakeholders to understand the risk profiles accurately.

In summary, ELBE and LGD-in-default are fundamental components in credit risk modeling for defaulted exposures. They inform the estimation of expected and unexpected losses, influencing both risk management strategies and regulatory capital requirements. Addressing the identified issues in modeling practices is crucial for improving the accuracy of loss estimations and strengthening the overall resilience of financial institutions against credit risk.

9.2 ELBE Calibration

The aim of the validation tool outlined in this section is to monitor the performance of Expected Loss Best Estimate (ELBE) models in terms of predictive ability, also known as calibration. The analysis of predictive ability focuses on ensuring that the ELBE parameter adequately predicts the loss rate in the event of a default—that is, ELBE values should constitute reliable forecasts of realised loss rates.

Currently, it is common practice to use accounting provisions as ELBE estimates, accounting for approximately 26% of models. However, the guidelines specify that institutions should estimate ELBE based on a Loss Given Default (LGD) model for non-defaulted exposures, calibrated to current economic conditions and taking into account all relevant post-default information. Consequently, it will no longer be permissible to assess ELBE on the basis of accounting provisions unless these provisions stem from a model that complies with the specified conditions.

Although it is not possible to accurately assess for all survey responses whether a model change will be necessary, it is expected that around 63% of ELBE models will need to be changed to comply with this policy choice. Institutions should therefore review their current ELBE estimation approaches and ensure alignment with the new guidelines.

To avoid excessive use of adjustments and hence increased subjectivity of resulting estimates, the guidelines specify that adjustments should be applied only where necessary. An adjustment may be necessary when the observed sensitivity of realised LGDs to economic conditions is not reflected in the ELBE estimates, either through direct use of macroeconomic factors in the model or through risk drivers that are sensitive to economic conditions. In all other cases, it is considered that the ELBE estimates based on the long-run average LGD for defaulted exposures sufficiently reflect current economic conditions, and no further adjustments are required.

Where an adjustment is applied, it should be adequately documented, including its rationale and calculation. This ensures transparency and allows for effective validation and regulatory assessment of the ELBE models.

ELBE Backtesting (t-test) The purpose of ELBE back-testing using a one-sample t-test for paired observations is to evaluate the predictive accuracy of the Expected Loss Best Estimate (ELBE) at various aggregation levels—such as portfolio, grade, pool, or segment—and at different reference points during default. This validation tool assesses whether there is a statistically significant difference between the ELBE estimates and the realised Loss Given Default (LGD) values.

In this approach, the one-sample t-test compares the ELBE estimates with the realised LGD values under the null hypothesis that the mean difference between ELBE and realised LGD is zero. This is a two-sided hypothesis test that assumes independent observations. Under the null hypothesis, the test statistic follows a Student's t-distribution with degrees of freedom equal to one less than the number of facilities included in the back-testing sample.

To determine statistical significance, the p-value is calculated by evaluating the cumulative distribution function of the Student's t-distribution at the absolute value of the test statistic. The p-value represents the probability of observing a test statistic as extreme as—or more extreme than—the one calculated, assuming that the null hypothesis is true. A low p-value indicates a statistically significant difference between the ELBE estimates and the realised LGD values, suggesting that the ELBE model may not be accurately predicting losses.

9.3 LGD-in-default Calibration

The calibration of LGD-in-default models is crucial to ensure that the estimated Loss Given Default (LGD) values accurately predict the realized loss rates upon default. The aim of the validation tools outlined in this section is to monitor the performance of LGD-in-default models in terms of predictive ability, or calibration.

As regulatory guidelines provide more prescriptive requirements for estimating LGD for defaulted exposures, institutions may need to adjust the calibration of some of their models. The analysis of predictive ability is aimed at ensuring that the LGD-in-default parameters adequately predict the loss rate in the event of a default—i.e., that LGD-in-default estimates constitute reliable forecasts of realized loss rates.

Calibration involves aligning the LGD estimates with the *long-run average LGD* calculated for each grade or pool. Institutions should provide additional calibration tests at the level of the relevant calibration segment to ensure that LGD estimates are consistent with observed loss experiences and reflect the risk characteristics of the exposures.

In practice, institutions may observe extremely high values of realized LGDs, often exceeding 100%, particularly for exposures with small outstanding amounts at the moment of default. In such cases, it is important to identify relevant risk drivers to differentiate these observations and adequately reflect these specific characteristics in the assignment to grades or pools. For institutions using a continuous rating scale in the LGD estimation, they may consider creating a separate calibration segment for such exposures.

By carefully calibrating LGD-in-default models and conducting thorough validation, institutions can ensure that their LGD estimates provide reliable predictions of loss rates, supporting effective risk management and regulatory compliance.

LGD-in-default Back-testing (t-test)

The objective of LGD-in-default back-testing using a one-sample t-test for paired observations is to assess the predictive ability of estimated loss given default (LGD) values at the portfolio level, as well as at the grade, pool, or segment level, at various points during default. This statistical tool compares the estimated LGD-in-default with the realised LGD under the null hypothesis that the estimated LGD-in-default is greater than the realised LGD.

Assuming independent observations, the one-sample t-test applies a one-sided hypothesis test. Under the null hypothesis, the test statistic follows an asymptotic Student's t-distribution with degrees of freedom equal to one less than the number of facilities in the back-testing sample. This means that if there are N facilities, the degrees of freedom would be $N - 1$.

To calculate the p-value for the test, the cumulative distribution function of the Student's t-distribution is evaluated at the test statistic. Specifically, the p-value is determined by subtracting this value from one, which reflects the probability of observing a test statistic at least as extreme as the one calculated, assuming that the null hypothesis is true.

By conducting this t-test, institutions can evaluate whether their LGD-in-default estimates are conservative and reflect the realised losses accurately. This assessment is crucial for regulatory compliance and for ensuring that the models used for estimating LGD-in-default are robust and reliable.

10 Benchmarking, Sensitivity, Stress Testing

Benchmarking, sensitivity analysis, and stress testing are critical validation activities in financial institutions. These processes ensure that models are robust, reliable, and accurately reflect the risks they are designed to measure.

Benchmarking involves comparing model outputs against established standards, alternative models, or external data sources to assess performance and identify discrepancies. Benchmarking analyses should be performed during the initial validation and at an adequate frequency thereafter, but at least every three years. For initial validations, the validation function can take into account the benchmarking analyses performed by the Credit Risk Control Unit (CRCU). Regular benchmarking helps institutions validate the accuracy of their models and ensures consistency with industry practices and regulatory requirements.

Sensitivity analysis examines how changes in input parameters or assumptions affect model outputs. This analysis is essential for understanding the model's responsiveness to variations and identifying parameters that significantly impact results. Sensitivity analysis should be conducted when a new pricing method is introduced into Value at Risk (VaR) or Stressed Value at Risk (sVaR) calculations, especially if the pricing method differs from that used for economic Profit and Loss (P&L) purposes. Additionally, regular validation should be performed to ensure that the impact of different pricing methods remains minimal. Institutions should develop a work plan to mitigate risks or improve the quality of any pricing functions or methods deemed inadequate according to internal assessments, such as scorecards indicating a red indicator.

Stress testing evaluates model performance under extreme but plausible scenarios. It helps institutions understand how models behave under adverse conditions and supports risk management decisions. A thorough review of all procedures applied to the data used for model development is essential. This includes data collection, data cleansing, data processing (e.g., normalization, treatment of collinearity), and data estimation (e.g., cash flow projections for specialized lending). A good practice is to complement this review with back-testing comparisons between these estimations—including projections extending beyond the one-year time horizon—and the subsequently realized values through out-of-time (OOT) validation tests.

Many institutions are currently implementing comprehensive independent validation frameworks across risk types. However, due to technical specificities, back-testing and the benchmarking of Internal Models Method (IMM) pricing functions have often remained the responsibility of model development teams. While internal model validation provides an independent challenge to underlying methodologies and reviews analyses of outcomes, some institutions have found that validation does not offer sufficient challenge when it operates under the purview of model development. To address this, institutions should ensure that validation functions are truly independent and provide robust challenges to model development activities, including back-testing and benchmarking.

Integrating benchmarking, sensitivity analysis, and stress testing into the validation process strengthens the overall risk management framework. These activities help institutions identify and address model weaknesses, ensure compliance with regulatory standards, and enhance the reliability of risk assessments.

10.1 Benchmarking Models

Benchmarking models is a fundamental process in model validation, aimed at ensuring that the models used for regulatory compliance and financial risk assessment are robust, accurate, and fit for purpose. This process involves comparing a model's performance against established standards or reference models to identify areas of improvement and confirm adherence to regulatory requirements.

1. **Data Collection and Selection Process**
   - A detailed description of the data collection and selection process is essential. This process should culminate in the creation of the *validation data set*, which encompasses all data sets used for the purpose of validation.
   - The data should be representative of the various types of exposures and sufficient in both quality and quantity to support rigorous validation activities.
2. **Definition of Default**
   - To ensure the default definition is specified correctly and applied consistently, documentation must include a comprehensive description of how default indications are operationalised.
   - This includes outlining the processes, sources of information, and responsibilities involved in identifying indications of default.
   - Both automatic mechanisms and manual processes should be described to demonstrate a thorough approach to default identification.
3. **Model Life Cycle**
   - The model life cycle encompasses several critical stages, each requiring meticulous attention:
     (a) *Development*: Involves the initial creation of the model, including the preparation of data and selection of appropriate methodologies.
     (b) *Calibration*: Adjusts model parameters to align with observed data, ensuring accurate predictive capabilities.
     (c) *Validation*: Assesses the model's performance through back-testing, sensitivity analysis, and other validation techniques to verify its reliability.
     (d) *Supervisory Approval*: Obtaining necessary approvals from regulatory bodies, if required, to ensure compliance with supervisory expectations.
     (e) *Implementation in Internal Processes*: Integrates the model into the organization's operational framework, affecting decision-making and risk management practices.
     (f) *Application and Review of Estimates*: Continuous monitoring and periodic review of the model's estimates to maintain accuracy over time.
4. **Project Reporting**
   - Comprehensive documentation, often in the form of a *project report*, is vital throughout the benchmarking process.

- The report should detail all aspects of the model's development, validation, and implementation, serving as a transparent record for internal and external stakeholders.

By systematically following these steps, organizations can ensure their models meet both internal standards and regulatory requirements. Benchmarking not only validates the model's effectiveness but also enhances its credibility, ultimately contributing to more informed decision-making in financial risk management.

Selecting Benchmarks Selecting appropriate benchmarks is a critical step in assessing the performance of internal models and ensuring regulatory compliance in finance. Benchmarks provide reference points that enable institutions to evaluate the impact of new methodologies on capital requirements. Since capital ratios are fundamental measures of financial strength, accurate benchmarking supports tools like stress testing by supplying correct starting points for vital risk parameters through regular benchmarking exercises at the European level.

When introducing benchmarking exercises, it is essential to ensure that these supervisory tools do not impede the adoption of new best practices. Competent authorities must make decisions regarding corrective actions that uphold the objectives of internal approaches and avoid the following pitfalls:

- Leading to *standardization or preferred methods*;
- Creating *wrong incentives*;
- Causing *herd behavior* among institutions.

The process of selecting benchmarks begins with the preparation and quality verification of an appropriate data set. It is important to recognize that the scope of data and the assessment of representativeness vary depending on the phase:

- **Model Development Phase**: Data is selected from a sample that is highly representative of the application portfolio. This sample should contain information on all relevant risk drivers to provide the best basis for effective risk differentiation.
- **Calibration Phase**: Data includes all observations from the relevant historical observation period to calculate long-run average default rates or Loss Given Default (LGD). Even if the data lacks sufficient representativeness, it should not be excluded. Instead, any issues should be assessed for their influence on risk quantification, and biases should be addressed through appropriate adjustments and the application of a margin of conservatism (MoC).

In the context of data scarcity, institutions should consider utilizing other relevant external data sources where available. For instance, external ratings can serve as a *challenger* to internal models. While comparisons with external ratings should not be used as objective benchmarks, they are valuable tools for identifying potential weaknesses and ensuring that all relevant information has been considered.

To effectively assess the performance of models under data limitations, best practices include:

- Employing external data sources as a means of challenging and validating internal models.
- Monitoring model specifications and tracking changes over time to capture shifts in performance (see Interaction Box [2]).
- Using metrics that measure the pure performance of the model, supplementing traditional validation techniques.

By selecting benchmarks that are compatible with new methodologies and practices, institutions can enhance model robustness without introducing unintended consequences. This approach ensures that benchmarking exercises contribute positively to capital assessment and risk management, while avoiding standardization, misaligned incentives, and herd behavior.

Internal and External Benchmarks In the context of model validation, benchmarking analyses play a crucial role in assessing the performance and robustness of internal models used in finance. Benchmarks can be classified into two categories: **internal benchmarks** and **external benchmarks**.

Internal benchmarks involve cross-checks within the institution's own data and models. Cross-checks should be carried out between different databases, when available, or between different providers. This practice is a sign of consistency and robustness. By comparing model outputs across various internal datasets or alternative internal models, institutions can identify discrepancies and ensure that the model performs reliably within different internal contexts.

External benchmarks entail the use of external data sources or ratings to challenge the internal model. Where external data are used, their representativeness, appropriateness, and consistency with regard to the institution should be assessed. It is a best practice to use external ratings as a challenger when a sufficient number of such ratings is available. However, the comparison with external ratings should not be used as an objective benchmark to assess the performance of the internal model. Instead, it should serve as a tool to search for potential weaknesses in terms of the model's effectiveness in considering all relevant information.

According to *Principle 4 – Performance Assessment*, even if the rating system has been developed using external data, the quantitative evaluation of its performance is expected to be performed first on the internal data. In cases where external data is used to circumvent data scarcity issues, the performance assessment of the rating system based on internal data can be complemented by an assessment using all data available. Challenges associated with the scarcity of data are further described in the section on validation in the context of data scarcity.

In summary, employing both internal and external benchmarks enhances the validation process by ensuring that internal models are consistent within the institution and robust against external standards. Careful consideration must be given to the suitability of external data, and the primary focus should remain on internal performance evaluation, leveraging external benchmarks as a supplementary tool to identify and address potential weaknesses.

Benchmarking Methods Benchmarking is a vital process in evaluating and validating the internal models employed by financial institutions. It involves comparing an institution's model outputs with industry benchmarks or results from peer institutions to assess performance and identify areas for improvement. One of the key objectives of benchmarking exercises is to provide tools to assess the effect of new methodologies on capital. Since capital ratios are the core measure of financial strength, tools such as stress testing can greatly benefit from accurate starting points for important risk parameters provided by regular benchmarking exercises at the European level.

However, while benchmarking provides significant insights, it is essential that these supervisory tools do not hinder the introduction of new best practices. Competent authorities must ensure that their decisions on the appropriateness of corrective actions maintain the objectives of an internal approach and therefore must not:

- Lead to standardisation or preferred methods;
- Create wrong incentives;
- Cause herd behaviour.

In conducting benchmarking analyses, institutions should consider methodologies that are compatible with the introduction of new methodologies and practices. Best practices in the context of data scarcity include:

- Utilizing robust validation activities and tests tailored to the specific constraints of the data;
- Comparing model outcomes after eliminating divergences arising from portfolio effects to enhance comparability between institutions;
- Disclosing benchmarking results to users to provide supplementary comparison tools, thereby enhancing trust in the internal models.

Such approaches enable institutions to assess the performance of their models effectively, even when data is limited. By facilitating comparisons of model outcomes across different institutions, benchmarking enhances transparency and helps maintain confidence in internal models without discouraging innovation or the adoption of new methodologies.

10.2 Sensitivity Analysis

Sensitivity analysis is a critical component in the validation of internal financial models. It involves systematically varying key model inputs and parameters to assess the impact on the model's outcomes. This process helps institutions understand the robustness of their models and identify the factors that most significantly influence risk assessments.

In accordance with Article 376(3)(b) of the Capital Requirements Regulation (CRR), institutions are required to perform sensitivity and scenario analyses to evaluate the qualitative and quantitative reasonableness of their internal models. The European Central Bank (ECB) considers it best practice for institutions to incorporate these analyses as part of their independent reviews and during initial and periodic validations.

Key aspects of sensitivity analysis include:

- **Assessment of Model Drivers:** Analyzing the main drivers of the Incremental Risk Charge (IRC) model, such as Probability of Default (PD) and Recovery Rates (RR), to understand how changes in these parameters affect the IRC and default risk measurements.
- **Data Adjustment:** Evaluating the extent to which historical data can be adjusted to align with materiality thresholds for calibrating risk estimates, ensuring that the data remains relevant and accurate for current risk assessments.
- **Analysis Documentation:** Providing a comprehensive list of analyses performed, including descriptions of their purposes, potential limitations, scopes of application, and the expected frequency of execution.

Institutions may be required by the ECB, based on Article 10 of the Single Supervisory Mechanism (SSM) Regulation, to submit the results of their sensitivity analyses. This allows the ECB to assess the appropriateness of the analyses performed and validate the reasonableness of the institutions' internal models.

By conducting thorough sensitivity analyses, institutions can identify vulnerabilities within their models and make necessary adjustments. This process not only enhances model accuracy but also strengthens overall risk management practices by ensuring that models remain reliable under varying conditions and assumptions.

10.3 Stress Testing

Stress testing is a critical risk management tool used by financial institutions to assess the potential impact of severe, yet plausible, adverse scenarios on their financial condition. The purpose of stress testing is to ensure that institutions maintain adequate capital and liquidity levels under stressed conditions, thereby safeguarding their solvency and the stability of the financial system.

The methodology of stress tests involves several key components:

- **Documentation**: A comprehensive documentation of the stress testing methodology is essential. This includes detailing the internal and external data sources used, as well as any expert judgment inputs. The documentation should be sufficiently detailed to allow third parties to understand the rationale for the chosen scenarios and to replicate the stress test.
- **Scenario Building**: Developing stress scenarios requires careful consideration of their severity, duration, and likelihood of occurrence. The methodology should outline how these scenarios are constructed, ensuring they are meaningful and reasonably conservative. This includes capturing severe but plausible recession scenarios

that could significantly impact the institution's credit risk profile and total capital requirements.

- **Impact Projection**: The stress testing process must include a methodology for projecting the impact of each scenario on relevant risk parameters. This involves assessing how different stress conditions would affect credit losses, asset valuations, liquidity positions, and capital adequacy.

- **Verification Steps**: Competent authorities are tasked with verifying that the stress tests include essential steps to ensure their effectiveness. These steps encompass the soundness of the scenario design, the appropriateness of the risk models used, and the accuracy of the impact assessments.

To be effective, stress tests should meet the following criteria:

- *Meaningful*: The tests must be relevant to the institution's specific risk profile and operational context, capturing the most significant vulnerabilities.

- *Reasonably Conservative*: Assumptions and parameters used in the stress tests should avoid underestimating potential risks, ensuring that the institution remains resilient under adverse conditions.

- *Transparent and Replicable*: The methodology and results should be transparent, allowing for replication and validation by third parties, including regulators and stakeholders.

By adhering to these principles, stress testing enables institutions to identify potential weaknesses in their risk management frameworks and to take proactive measures to mitigate adverse effects. This not only enhances the institution's own resilience but also contributes to the overall stability of the financial system.

11 Advanced Topics

The field of credit risk model validation has evolved significantly, presenting new challenges and complexities for financial institutions. This section explores advanced and specialized topics that are critical for ensuring robust and compliant credit risk models.

One of the most pressing areas is the validation and governance of **counterparty credit risk (CCR)** models. Supervisory reviews have identified a high number of significant findings in this area, with all CCR investigations uncovering at least one issue and 60% of cases involving high-severity findings. Key modeling topics where deficiencies have been noted include:

- **Trade Coverage**: Ensuring comprehensive inclusion of all relevant trades within the model scope.
- **Margin Period of Risk (MPR)**: Accurately determining the time horizon over which a bank is exposed to counterparty default risk.
- **Collateral Management**: Properly accounting for collateral agreements and their impact on exposure calculations.
- **Initial Margin**: Incorporating initial margin requirements into risk assessments to reflect true exposure levels.
- **Risk Factors and Calibration**: Selecting appropriate risk factors and calibrating models to current market conditions.

To address these challenges, institutions have developed validation tools that focus on the quantitative aspects of internal models. Providing detailed instructions and clear methodologies is essential to facilitate implementation and reduce ambiguity. However, despite standardized regulatory requirements, variability in validation processes across institutions limits the ability to perform meaningful comparisons.

Non-model-specific aspects also play a crucial role in credit risk model validation. Findings have highlighted issues related to the organization and activities of the **internal validation function**, as well as processes surrounding **roll-out and permanent partial use (PPU)** and **model change management**. All institutions reviewed received feedback indicating that certain practices did not align with regulatory expectations. Specifically:

- **Internal Validation Function**: Ensuring independence, adequate resources, and clear methodologies within validation teams.
- **Roll-Out and PPU**: Managing the phased implementation of internal models and the use of standardized approaches for certain portfolios.
- **Model Change Management**: Establishing robust procedures for model updates, including documentation and approval processes.

Regulatory bodies have responded by issuing feedback letters with recommendations and, in some cases, supervisory decisions imposing obligations to address identified deviations. This underscores the importance of aligning practices with the regulatory framework to mitigate compliance risks.

While these topics have been primarily considered in relation to credit risk models, they are also pertinent to **market risk** and **CCR** models. For these risk types, critical sub-topics such as internal validation are assessed through targeted on-site investigations, emphasizing the need for consistent validation practices across all risk areas.

Financial institutions must develop robust validation processes to assess their internal models effectively. Despite uniform regulatory requirements, the diversity in models and practices poses challenges for comparability and benchmarking. To enhance consistency, institutions should focus on:

- Implementing detailed validation methodologies with clear instructions.
- Enhancing transparency and documentation to facilitate supervisory review.
- Promoting best practices through industry collaboration and guidance.

In conclusion, addressing advanced topics in credit risk model validation is essential for maintaining the integrity and reliability of internal models. By prioritizing areas with significant findings and fostering strong validation frameworks, institutions can better manage credit risk and meet regulatory expectations.

11.1 Low Default Portfolios

Low Default Portfolios (LDPs) present specific challenges in risk modelling and validation due to the scarcity of historical default data. The small number of defaults in these portfolios makes reliable statistical modelling difficult. Consequently, expert judgement and the individual bank's experience play a more significant role for LDPs than for other portfolios.

An appropriate solution for the definition and treatment of LDPs involves acknowledging these limitations and adjusting modelling approaches accordingly. Key considerations include:

- **Estimation of Risk Parameters**: With limited default data, traditional statistical methods for estimating probabilities of default (PD), loss given default (LGD), and exposure at default (EAD) become less reliable. Institutions may need to employ alternative techniques, such as using proxy data, incorporating external data sources, or applying conservative estimates to compensate for the lack of empirical evidence.
- **Role of Expert Judgement**: Expert judgement becomes crucial in the assessment of LDPs. Banks should establish robust frameworks for integrating expert opinions, ensuring that such inputs are systematic, transparent, and subject to validation.

- **Model Validation and Monitoring**: The usual back-testing and benchmarking methods are less effective for LDPs due to the insufficient number of observed defaults. Validators should focus on qualitative aspects, stress testing, sensitivity analysis, and scenario analysis to evaluate model performance. Regular monitoring is essential to capture any changes in portfolio characteristics or external conditions.
- **Data Quality and Maintenance**: Maintaining high-quality data is vital for LDPs. Institutions must ensure accurate implementation of the definition of default and meticulous data maintenance practices. Even with few defaults, each data point carries significant weight in the modelling process.

Figure 8 shows the distribution of shortcomings per data quality topic and by severity. Similar to the data quality review of models for retail and SME portfolios, the findings for LDP models were either categorised under the institution's technical implementation of the definition of default or in relation to data maintenance.

To address these challenges, institutions should consider the following modelling restrictions and treatments for LDPs:

- **Conservative Modelling Approaches**: Adopt conservative assumptions in model development to account for uncertainties arising from limited data.
- **Use of External Data**: Where appropriate, supplement internal data with external sources, such as industry-wide default statistics, to enhance model robustness.
- **Regular Validation Adjustments**: Increase the frequency of model validations and updates to promptly reflect any new information or changes in the portfolio.
- **Enhanced Governance**: Strengthen model governance frameworks to oversee the use of expert judgement and ensure compliance with regulatory standards.

In conclusion, the effective management of LDPs requires a balanced approach that mitigates the limitations of statistical modelling through the judicious use of expert judgement and adherence to rigorous data quality practices. By implementing appropriate solutions and modelling restrictions, banks can enhance the reliability of their risk assessments for low default portfolios.

11.2 Overfitting, Model Selection, Data

Machine learning (ML) models have become integral in financial risk management and regulatory compliance. However, they are susceptible to challenges such as overfitting, improper model selection, and data quality issues. This section discusses the importance of addressing these challenges to enhance model performance and reliability.

Avoiding Overfitting

Overfitting occurs when a model is excessively tailored to the training data, capturing noise instead of the underlying patterns. This leads to high performance on the development sample but poor generalization to new, unseen data. In the financial context, overfitting can result in inaccurate risk predictions and misguided strategic decisions.

To mitigate overfitting:

- *Model Comparison*: Institutions should compare model performance across different datasets, including validation and test samples, to ensure consistent results.
- *Bias Detection*: Implement techniques to detect potential biases that may arise from overfitting to the training data.
- *Regularization Methods*: Employ regularization techniques that penalize model complexity to prevent overfitting.
- *Cross-Validation*: Use cross-validation methods to evaluate model performance on multiple subsets of the data.

Model Selection and Hyperparameter Tuning

The choice of model architecture and hyperparameters significantly influences the model's ability to generalize. Hyperparameters, which define the model's structure and learning process, are often determined by human judgment or through optimization procedures.

Key considerations include:

- *Rationale Verification*: The validation team should scrutinize the reasoning behind selected hyperparameters to ensure they are justified and appropriate.
- *Avoiding Bias*: When hyperparameters are optimized by minimizing training error, it's crucial to ensure this process does not introduce bias or overfitting.
- *Expertise Requirement*: Complex models may require deep methodological knowledge to understand the implications of hyperparameter choices fully.
- *Robustness Checks*: Conduct sensitivity analyses to assess how changes in hyperparameters affect model performance.

Ensuring Data Quality and Representativeness

High-quality, representative data are essential for reliable model outcomes. Large volumes of data do not guarantee accuracy if the data are flawed or unrepresentative of the application portfolio.

Institutions should:

- *Assess Data Quality*: Ensure data are accurate, complete, and appropriate for the intended use, especially when dealing with unstructured data.
- *Evaluate Representativeness*: Carefully assess whether external data sources reflect the characteristics of the institution's portfolio.
- *Monitor Performance Impact*: Verify if reduced data representativeness leads to decreased model performance on internal customer segments.

- *Data Governance*: Implement robust data governance frameworks to manage data quality and lineage.

Challenges in Model Understanding and Implementation

The use of advanced ML techniques introduces additional complexities:

- *Mathematical Assumptions*: Complex models rely on mathematical hypotheses that may be challenging to validate or understand fully.
- *Implementation Difficulty*: Such models often require significant computational resources and specialized expertise.
- *Partial Understanding*: No single technique provides a complete understanding of the model; different methods offer partial insights that vary in usefulness depending on the context.

Institutions must recognize these challenges and allocate sufficient resources to address them effectively.

Conclusion

Addressing overfitting, making informed model selections, and ensuring data integrity are critical for the success of ML models in finance. By placing particular attention on these areas, institutions can enhance model performance, reduce risks associated with model bias, and ensure compliance with regulatory standards. Continuous validation and careful consideration of model design choices contribute to building robust models that perform reliably across different scenarios.

11.3 Machine Learning Models

Machine Learning (ML) models have increasingly been adopted in the finance industry for various applications, including credit risk assessment. In the context of credit risk, ML models offer advanced techniques for modeling and predicting key risk parameters such as Probability of Default (PD), Loss Given Default (LGD), Expected Loss Best Estimate (ELBE), Exposure at Default (EAD), and Credit Conversion Factor (CCF). These models can capture complex nonlinear relationships and interactions between variables, which traditional statistical models may not fully address.

According to the IIF 2019 report, the most common use of ML within credit risk is in the areas of credit decisions and pricing, followed by credit monitoring, collections, restructuring, and recovery. In contrast, the use of ML is avoided for regulatory areas such as capital requirements for credit risk, stress testing, and provisioning. Regulatory requirements are perceived as a challenge for the application of ML models in these areas, as these models are more complex to interpret and explain.

Despite these challenges, institutions are exploring the integration of ML models within the Internal Ratings-Based (IRB) framework. The types of ML models and algorithms that are currently used or planned for use include:

- **Random Forests**: An ensemble learning method that constructs multiple decision trees and outputs the mode of their predictions.
- **k-Nearest Neighbors (k-NN)**: A non-parametric method used for classification and regression by analyzing the k closest training examples in the feature space.
- **Support Vector Machines (SVM)**: A supervised learning model that analyzes data for classification and regression analysis.
- **Neural Networks**: Computational models inspired by the human brain, capable of modeling complex patterns and prediction problems.
- **Gradient Boosting Machines**: An ensemble technique that builds models in a stage-wise fashion and generalizes them by allowing optimization of an arbitrary differentiable loss function.

These ML models are being considered for the estimation of key parameters in credit risk modeling, such as PD, LGD, ELBE, EAD, and CCF. By leveraging ML techniques, institutions aim to enhance predictive accuracy and gain deeper insights into credit risk.

Several learning paradigms may be used to train these ML models, including supervised learning, unsupervised learning, and semi-supervised learning. Each paradigm offers distinct approaches depending on the availability of labeled data and the specific objectives of the modeling exercise. A comprehensive categorization of these learning paradigms is provided in Section 1.3 of the Report on Big Data and Advanced Analytics (BD&AA).

However, the adoption of ML models in the IRB context is not without challenges. Concerns over interpretability, regulatory compliance, and the cost of implementation may deter institutions from fully embracing these advanced techniques. Regulatory requirements demand that models are transparent and explainable, which can be difficult to achieve with certain ML models. Institutions must weigh the potential benefits against these challenges when deciding whether to integrate ML models into their credit risk assessment processes.

Considerations for Implementing ML Models in Credit Risk:

- *Interpretability*: Ensuring that the ML models provide explanations for their predictions to satisfy regulatory standards.
- *Regulatory Compliance*: Aligning ML model usage with regulatory requirements for risk modeling and validation.
- *Cost*: Evaluating the costs associated with developing, implementing, and maintaining ML models.
- *Model Validation*: Establishing robust processes for validating ML models to ensure their reliability and effectiveness.

In conclusion, while ML models offer significant potential for enhancing credit risk modeling, institutions must carefully consider the implications of their use within the IRB framework. Balancing the advantages of advanced analytics with the need for interpretability and compliance is crucial for the successful integration of ML models in regulatory credit risk assessment.

11.4 Explainable AI (XAI)

In the realm of credit risk modeling, the adoption of Machine Learning (ML) techniques presents both opportunities and challenges for financial institutions. While ML models offer enhanced predictive power and have been utilized in internal credit approval processes, their integration into Internal Ratings-Based (IRB) models for calculating regulatory capital requirements has been slow. One of the primary obstacles is the complexity of advanced ML models, which leads to difficulties in interpreting results, ensuring management's understanding, and justifying outcomes to supervisory authorities.

Explainable AI (XAI) emerges as a critical solution to these challenges by enhancing the transparency and interpretability of ML models. XAI techniques enable institutions to:

- **Interpret Model Results:** By providing insights into how individual variables contribute to the predictions, XAI helps demystify complex algorithms, making the results more accessible to stakeholders.
- **Ensure Management Understanding:** Clear explanations of model behavior facilitate better comprehension among management functions, supporting informed decision-making and effective risk management.
- **Justify Outcomes to Supervisors:** Regulatory bodies require that institutions can substantiate their models' decisions. XAI aids in meeting these demands by offering transparent justifications aligned with regulatory expectations.

The importance of XAI is further underscored by evolving regulatory landscapes. The European Union's proposed legislative framework on artificial intelligence (the *AI Act*) includes provisions classifying the use of AI for evaluating creditworthiness or establishing credit scores as high-risk use cases. This emphasis on transparency and accountability signifies a regulatory push towards models that are not only accurate but also explainable.

Financial institutions are adopting various strategies to enhance the explainability of their ML models, such as:

- **Ex Post Analysis Tools:** Utilizing post-hoc interpretability methods to analyze and describe the influence of individual variables on model predictions after the model has been trained.
- **Algorithmic Constraints:** Introducing constraints during the modeling process to reduce complexity, which can lead to more interpretable models without significantly sacrificing performance.

By prioritizing explainability, institutions can bridge the gap between advanced ML capabilities and the stringent demands of credit risk regulation. This alignment not only facilitates compliance but also promotes trust among stakeholders by ensuring that credit decisions are transparent and justifiable.

In conclusion, Explainable AI plays a pivotal role in modernizing credit risk modeling. It enables institutions to leverage sophisticated ML models while maintaining the necessary transparency and accountability required in the financial industry. As regulatory expectations evolve and the use of AI becomes more prevalent, the focus on XAI will continue to grow, making it an essential component of effective and compliant credit risk management.

11.5 Changing Economic Environment

In the dynamic landscape of finance, economic conditions are subject to continual change, influencing the performance and reliability of risk models. It is paramount for institutions to adapt their models to reflect these evolving conditions to maintain accuracy and regulatory compliance.

Models exhibit varying degrees of sensitivity to economic fluctuations:

- For approximately 3% of the models, the rating assignment process is described as *fully sensitive* to economic conditions.
- Approximately 26% of the models are considered *highly sensitive* to economic conditions.
- About 33% of the models have a *low sensitivity* to economic changes.

These statistics highlight the necessity for regular model evaluations and recalibrations. A recalibration is generally required when there is a significant shift in economic conditions, changes in the institution's processes, or alterations in the underlying data. Such proactive adjustments are essential to prevent issues like permanent overfitting, where a model might perform well on historical data but fail under new economic scenarios.

Regulatory guidelines underscore the importance of this adaptability. As stipulated:

> Institutions should specify conditions under which the analyses referred to in paragraph 218 should be performed more frequently than annually, such as major changes in the risk profile of the institution, credit policies or relevant IT systems. Institutions should perform the review of the PD or LGD model whenever they observe significant change in economic conditions as compared with the economic conditions underlying the dataset used for the purpose of model development.

Even when potential model changes have minor impacts, institutions must assess whether adaptation during retraining is necessary. This ensures models remain robust and accurately reflect the current economic environment.

In conclusion, adapting models to changing economic conditions is vital for effective risk management. Institutions must remain vigilant and responsive to economic shifts to maintain the integrity and reliability of their risk assessment models.

11.6 Specialised Lending Exposures

Specialised lending exposures refer to types of lending where the repayment is highly dependent on the project's cash flows rather than on the financial position of the borrower.

Due to the unique risk characteristics of these exposures, banks employ the Slotting approach for risk assessment and capital computation.

The Slotting approach categorises exposures into predefined slots based on specific risk characteristics. Each slot corresponds to a particular risk weight and expected loss (EL) estimate. The adequacy of slot assignment is crucial, as it directly influences both the risk weights and the expected loss associated with the exposure.

The adequacy of slot assignment is assessed using the following areas of investigation:

- **Evaluation of Risk Characteristics**: Assessing the exposure's attributes to ensure correct slot allocation.
- **Consistency in Slotting Criteria**: Maintaining uniform application of slotting criteria over time.
- **Data Integrity**: Ensuring the reliability and accuracy of data used in the slotting process.
- **Regulatory Compliance**: Aligning slotting practices with regulatory requirements.

The validation tools outlined in this section aim to ensure the adequacy of slot assignment. At the slot level, the expected loss (EL) for a given slot is calculated pursuant to Article 158(6) of the CRR for that slot. The data basis is restricted to customers who would have been assigned to a specific slot at the beginning of the year in which the customer defaulted, based on the slotting method in place at the end of the relevant observation period. The historical default rates calculated correspond to the specific slot in question.

Stability is a key consideration in the slotting process. Monitoring the stability of slot assignments over time ensures that the Slotting approach remains effective and reflective of the underlying risk.

Guidance provided in the Final draft EBA/RTS/2016/02 published on 13 June 2016 should be considered when implementing or validating the slotting process. This document provides detailed instructions on the application of the Slotting approach and emphasizes the importance of adhering to regulatory standards.

12 Statistical Tables

In this section, we provide critical value tables for the t-distribution and chi-squared distribution. These tables are essential for statistical analyses in model validation and regulatory compliance within the finance industry.

The above tables provide critical values corresponding to various significance levels, which are commonly used in hypothesis testing. These values are instrumental when assessing the statistical significance of test results, such as evaluating the goodness-of-fit or comparing sample means during model validation processes.

Table 1: Critical Values of the t-Distribution

Degrees of Freedom	0.10	0.05	0.025	0.01	0.005
1	6.314	12.706	25.452	63.657	127.321
2	2.920	4.303	6.205	9.925	14.089
3	2.353	3.182	4.177	5.841	7.453
4	2.132	2.776	3.495	4.604	5.598
5	2.015	2.571	3.163	4.032	4.773
6	1.943	2.447	2.969	3.707	4.317
7	1.895	2.365	2.841	3.499	4.029
8	1.860	2.306	2.752	3.355	3.833
9	1.833	2.262	2.685	3.250	3.690
10	1.812	2.228	2.634	3.169	3.581

Table 2: Critical Values of the Chi-Squared Distribution

Degrees of Freedom	0.995	0.975	0.95	0.90	0.80
1	0.000	0.001	0.004	0.016	0.064
2	0.010	0.051	0.103	0.211	0.446
3	0.072	0.216	0.352	0.584	1.005
4	0.207	0.484	0.711	1.064	1.649
5	0.411	0.831	1.145	1.610	2.343
6	0.676	1.237	1.635	2.204	3.070
7	0.989	1.690	2.167	2.833	3.822
8	1.344	2.180	2.733	3.490	4.594
9	1.735	2.700	3.325	4.168	5.380
10	2.156	3.247	3.940	4.865	6.179

13 Glossary of Terms

General Topics Subjects that cover broad aspects of regulatory compliance and model validation in finance. These may include risk assessment methodologies, regulatory frameworks, and industry best practices.

Handbook Structure The organizational layout of the handbook, which is illustrated in the figure below and can be summarized as follows:

- Introduction to regulatory compliance and model validation.
- Detailed chapters on specific models and validation techniques.
- Appendices containing supplementary materials and references.

Definition of Default The specific criteria that determine when a financial obligation is considered to have defaulted. This includes quantitative thresholds and qualitative factors as outlined in the guidelines (GL). Clarifications on the application of this definition are provided to ensure consistent interpretation.

Sheet "Definitions" A dedicated section or spreadsheet within the handbook that lists and explains all key terms and abbreviations used throughout the document. This sheet serves as a quick reference for readers to understand technical language and concepts.

Calibration The process of adjusting model parameters to align the model's outputs with observed market data or empirical evidence. Calibration ensures that the model accurately reflects real-world conditions. Figure 2 provides an illustration of the notion of calibration.

Figure 2: Illustration of Notion of Calibration A visual representation included in the handbook that demonstrates how calibration is performed within a model. This figure aids in understanding the calibration process by providing a concrete example.

Guidelines (GL) Official documents issued by regulatory bodies that provide directives and standards for financial institutions. These guidelines inform the development and validation of financial models to ensure compliance with regulatory requirements.

Regulatory Compliance Adherence to laws, regulations, guidelines, and specifications relevant to financial institutions. Regulatory compliance involves implementing processes and controls to ensure that all organizational activities meet legal and regulatory standards.

Model Validation The set of processes and activities undertaken to verify that a financial model is performing as intended. Model validation assesses the model's conceptual soundness, performance, and ongoing appropriateness in the face of changing market conditions.

14 Code Library (Python/R)

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from scipy.stats import kstest

# Load the dataset
# Replace 'your_data.csv' with the path to your dataset
data = pd.read_csv('your_data.csv')

# Display the first few rows
print(data.head())

# Data quality checks
# Check for missing values
print("Missing values per column:")
print(data.isnull().sum())

# Drop rows with missing values (if applicable)
data = data.dropna()

# Check for duplicate rows
print("Number of duplicate rows:")
print(data.duplicated().sum())

# Drop duplicate rows
data = data.drop_duplicates()

# Generate descriptive statistics
print("Descriptive statistics:")
print(data.describe())

# Correlation matrix
correlation_matrix = data.corr()

# Visualize the correlation matrix
plt.figure(figsize=(12,8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix Heatmap')
plt.show()

# Histogram of variables
for column in data.select_dtypes(include=np.number).columns:
    plt.figure(figsize=(8,6))
    sns.histplot(data[column], kde=True)
    plt.title(f'Distribution of {column}')
    plt.xlabel(column)
    plt.ylabel('Frequency')
    plt.show()

# Scatter plots between variables
numeric_columns = data.select_dtypes(include=np.number).columns
for i in range(len(numeric_columns)):
```

```
    for j in range(i+1, len(numeric_columns)):
        plt.figure(figsize=(8,6))
        sns.scatterplot(x=data[numeric_columns[i]], y=data[
            numeric_columns[j]])
        plt.title(f'{numeric_columns[i]} vs {numeric_columns[j]}')
        plt.xlabel(numeric_columns[i])
        plt.ylabel(numeric_columns[j])
        plt.show()

# Regression analysis
# Define the dependent variable 'y' and independent variables 'X'
# Replace 'dependent_variable' and 'independent_variable_list' with
    your column names
y = data['dependent_variable']
X = data[['independent_variable1', 'independent_variable2']]

# Add a constant term for the intercept
X = sm.add_constant(X)

# Fit the regression model
model = sm.OLS(y, X).fit()

# Print regression results
print(model.summary())

# Residual analysis
residuals = model.resid

# Plot residuals histogram
plt.figure(figsize=(8,6))
sns.histplot(residuals, kde=True)
plt.title('Residuals Distribution')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.show()

# QQ plot of residuals
sm.qqplot(residuals, line='s')
plt.title('QQ Plot of Residuals')
plt.show()

# Kolmogorov-Smirnov test on residuals
ks_statistic, p_value = kstest(residuals, 'norm')
print(f"KS Statistic: {ks_statistic}")
print(f"P-value: {p_value}")

if p_value < 0.05:
    print("Residuals are not normally distributed.")
else:
    print("Residuals are normally distributed.")

# Time series analysis (if applicable)
# Ensure 'date_column' is of datetime type and set as index
# Replace 'date_column' with your date column name
data['date_column'] = pd.to_datetime(data['date_column'])
data.set_index('date_column', inplace=True)

# Plotting time series data
```

```
plt.figure(figsize=(12,6))
data['dependent_variable'].plot()
plt.title('Time Series Plot')
plt.xlabel('Date')
plt.ylabel('Dependent Variable')
plt.show()

# Autocorrelation plot
from pandas.plotting import autocorrelation_plot
plt.figure(figsize=(10,6))
autocorrelation_plot(data['dependent_variable'])
plt.title('Autocorrelation Plot')
plt.show()

# ARIMA model for time series forecasting
from statsmodels.tsa.arima.model import ARIMA

# Fit ARIMA model
model = ARIMA(data['dependent_variable'], order=(1,1,1))
model_fit = model.fit()

# Print model summary
print(model_fit.summary())

# Forecast future values
forecast = model_fit.forecast(steps=10)
print("Forecasted values:")
print(forecast)

# Model validation: Backtesting
# Split data into training and testing sets
train_size = int(len(data) * 0.8)
train, test = data[0:train_size], data[train_size:]

# Fit model on training data
model = ARIMA(train['dependent_variable'], order=(1,1,1))
model_fit = model.fit()

# Forecast on test data
forecast = model_fit.forecast(steps=len(test))
forecast = forecast.rename('Forecast')

# Plot actual vs forecasted values
plt.figure(figsize=(12,6))
plt.plot(test.index, test['dependent_variable'], label='Actual')
plt.plot(test.index, forecast, label='Forecast')
plt.title('Actual vs Forecasted Values')
plt.xlabel('Date')
plt.ylabel('Dependent Variable')
plt.legend()
plt.show()

# Calculate forecast errors
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(test['dependent_variable'], forecast)
print(f"Mean Squared Error: {mse}")
```

15 Case Studies

In this section, we present two case studies that illustrate common challenges and best practices in model validation and regulatory compliance within the financial industry. These case studies incorporate findings from general topic reviews and emphasize the importance of rigorous validation processes for financial models.

Case Study 1: Validation of a Credit Risk Model

A multinational bank developed an internal credit risk model to assess the probability of default (PD) for its corporate lending portfolio. The model utilized various financial indicators, macroeconomic variables, and qualitative factors to predict defaults.

During the validation process, several critical issues were identified:

- **Data Quality Issues**: The input data contained discrepancies and missing values, particularly in historical default records, which undermined the model's reliability.
- **Model Overfitting**: The model showed high accuracy on the training dataset but performed poorly on new, unseen data, indicating overfitting.
- **Regulatory Non-Compliance**: The model did not fully comply with regulatory requirements stipulated in Regulation 9, specifically regarding the incorporation of findings from general topics review.

To address these issues, the bank implemented the following actions:

- Enhanced data collection and preprocessing procedures to improve data quality and completeness.
- Simplified the model by reducing complexity, thereby improving its generalization capabilities on new data.
- Updated the model development process to include findings from the general topics review, ensuring compliance with Regulation 9.

This case highlights the importance of data integrity, model simplicity, and adherence to regulatory standards in developing robust credit risk models.

Case Study 2: Implementation of Regulatory Compliance in Market Risk Models

An investment firm developed multiple market risk models to estimate potential losses in its trading portfolio. These models incorporated various financial instruments and market factors to calculate Value at Risk (VaR).

Upon review, the following challenges were uncovered:

- **Inconsistent Model Outputs**: Different models produced conflicting risk estimates for similar market conditions, causing confusion in risk assessment.

- **Lack of Documentation**: There was insufficient documentation detailing the models' methodologies, assumptions, and limitations.
- **Non-Integration of Regulatory Findings**: The models did not include findings from the general topics review as required by regulatory standards, particularly those outlined in Section 2 of the compliance guidelines.

The firm responded by:

- Standardizing the modeling approaches to ensure consistency in risk assessments across different models.
- Developing comprehensive documentation for each model, enhancing transparency and understanding among stakeholders.
- Integrating findings from the general topics review into the models, achieving full compliance with Section 2 of the regulatory guidelines.

This case underscores the necessity of consistency, thorough documentation, and the integration of regulatory findings in the development and validation of market risk models.

These case studies demonstrate that meticulous model validation and a strong commitment to regulatory compliance are essential for financial institutions. By addressing data quality issues, preventing overfitting, ensuring consistency, and integrating regulatory findings, organizations can enhance the reliability of their models and maintain confidence among regulators and stakeholders.