End-to-End Credit Default Risk Modeling and Strategy Simulation

Credit Risk Analytics Framework

Nicolas Valderrama Barrera

Data Analytics / Credit Risk Portfolio Project

Domain:	Credit Risk Analytics
Scope:	PD Modeling, Calibration, Scorecards, Strategy Simulation
Stack:	AWS (S3, Glue, Athena, SageMaker), Python, Power BI
Data:	Inspired by Home Credit Default Risk (Kaggle)
Version:	1.0

This report is for educational and portfolio demonstration purposes only.

December 23, 2025

Contents

Executive Summary

This project presents an end-to-end credit risk analytics framework designed to estimate the Probability of Default (PD) for retail loan applicants and translate predictive outputs into actionable credit decisions.

Using a synthetic banking dataset inspired by the Home Credit Default Risk portfolio (Home Credit Group 2018), the solution spans the full analytics lifecycle: data ingestion, feature engineering, ETL, model development, calibration, score banding, and portfolio-level decision analysis. The objective is not only predictive and evaluate performance, but interpretability, calibration quality, and decision usability, which are essential in regulated credit environments.

A cloud-native architecture was implemented on AWS, with raw data stored in Amazon S3, data transformation handled through AWS Glue and PySpark, and curated datasets stored in partitioned Parquet format in S3. Amazon Athena was used for data validation and analytical queries, while model development and evaluation were performed in Amazon SageMaker (Amazon Web Services 2025). Final outputs were integrated into Power BI to enable executive-level portfolio monitoring and strategy simulation.

Different classification models were evaluated, including Logistic Regression, Support Vector Machines, and Random Forests. While non-linear models showed comparable discrimination performance, Logistic Regression was selected as the production model due to its interpretability, scalability and stability (Scikit-learn developers 2025).

The final model achieves an AUC of approximately 73%, with stable performance across training and validation datasets and strong recall of default events. Given that raw model probabilities are often miscalibrated, scaling methods was applied to ensure that predicted PDs closely match observed default rates. Calibration quality was validated using Brier Score and calibration curves, demonstrating a substantial improvement in probability accuracy without loss of discriminatory power. Compared to uncalibrated model outputs, the calibrated framework enables materially more reliable approval thresholds, reducing the risk of systematic under- or over-estimation of portfolio risk.

Calibrated PDs were converted into a monotonic scorecard and five risk bands (A–E), enabling transparent segmentation of applicants by risk level. The banding structure exhibits strong monotonicity, with observed default rates increasing consistently from Band A to Band E, supporting defensible credit policy design.

A credit strategy simulator was developed to evaluate approval policies under different risk appetites. For example, approving applicants in Bands A–C results in an approval rate of approximately 59.4%, while maintaining an average PD below 4% and significantly reducing expected defaults compared to broader approval strategies. This highlights the trade-off between portfolio growth and risk exposure.

1 Business Problem and Decision Context

A retail lender must balance two competing objectives: (1) grow the approved loan portfolio and (2) control credit losses. The decision is operationalized through an approval policy that determines which applicants are accepted or rejected at origination.

This problem is inspired by the Home Credit Default Risk use case, where consumer lenders aim to expand access to credit—often using alternative data sources—while still managing default risk. The objective is to better identify applicants who are likely to repay, so that creditworthy customers are not declined unnecessarily, and riskier approvals can be controlled through policy thresholds and portfolio monitoring.

The core deliverable is a calibrated PD score that can be translated into discrete risk bands (A–E). These bands support consistent credit decisioning (e.g., approve A–C), clear communication to business stakeholders, and monitoring of risk distribution over time. In addition, the framework quantifies trade-offs between approval rate and expected defaults, allowing risk managers to align originations with an explicit risk appetite.

2 Data and Architecture Overview

This section describes the data sources used in the project and the cloud-based architecture implemented to support scalable data processing, model development, and portfolio analytics. The design follows common industry practices in credit risk analytics, emphasizing data lineage, reproducibility, and separation between raw, curated, and analytical layers.

2.1 Data Sources

The modeling framework is built using a synthetic retail banking dataset inspired by the Home Credit Default Risk portfolio. The primary data sources include application-level information (demographics, income, employment), historical credit bureau records, and observed default outcomes.

For this project, the following datasets were used:

- **Application Train**: labeled loan applications with default outcomes.
- **Application Test**: unlabeled applications for out-of-sample evaluation.
- **Bureau Data**: external credit history aggregated at applicant level.

Altought additional auxiliary datasets are available in the original source, the selected tables provide sufficient coverage to construct a first view for PD model while keeping the pipeline focused and interpretable.

2.2 Cloud Architecture

A cloud-native architecture was implemented on Amazon Web Services (AWS) to support scalable data processing, model development, and analytics. The architecture follows a layered design commonly used in data-intensive financial applications.

The main architectural components and their roles are summarized below:

- **Amazon S3:** centralized storage for raw, curated, and model output data.
- **AWS Glue:** scalable ETL processing, feature engineering, and schema management.
- **Amazon Athena:** serverless querying and data validation on curated datasets.
- **Amazon SageMaker:** model development, training, and evaluation environment.

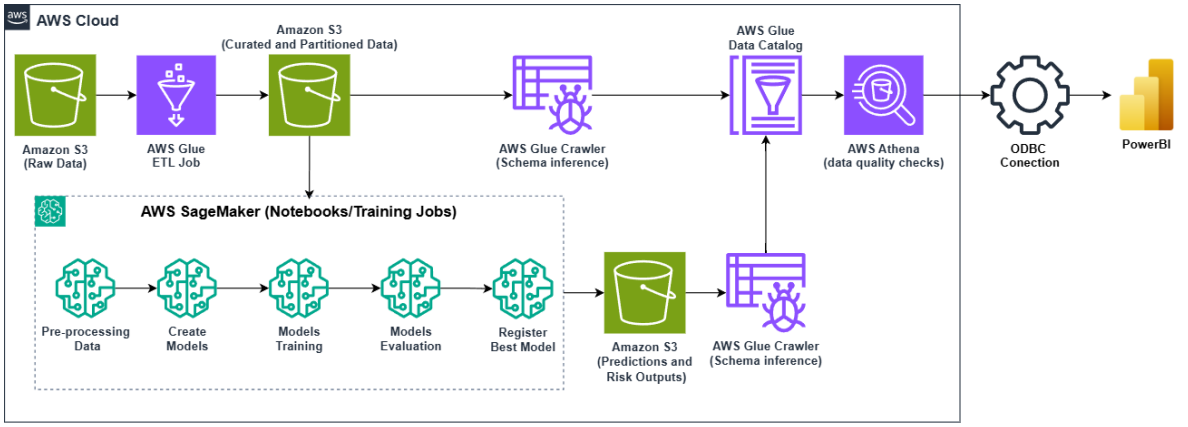

Figure 1: End-to-end cloud architecture for credit risk modeling, calibration, and strategy simulation.

- **Power BI:** visualization and decision-support layer for portfolio monitoring and credit strategy simulation.

Figure 1 illustrates the end-to-end data flow from raw data ingestion to business-facing analytics, highlighting the separation between data storage, processing, modeling, and decision support layers.

2.3 ETL and Data Governance

The ETL pipeline is implemented using AWS Glue and PySpark, transforming raw application and bureau data into analysis-ready datasets. Data cleaning steps include removal of highly correlated features, handling of missing values using domain-appropriate strategies (median imputation, zero-imputation for count variables, and explicit missing flags), and encoding of categorical variables.

Credit bureau data is aggregated at the applicant level using summary statistics such as counts of active and closed loans, exposure metrics, delinquency indicators, and temporal features related to credit history. These aggregated features are merged with application data to form a unified modeling dataset. Curated datasets are stored in Parquet format and partitioned by a technical identifier to optimize query performance and scalability.

This ETL design ensures reproducibility and traceability of features, supporting auditability and future model validation.

3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the structure of the credit portfolio, assess baseline risk characteristics. The analysis focuses on portfolio-level patterns rather than exhaustive variable-by-variable exploration.

A comprehensive feature-level EDA was performed during data preparation to inform feature selection, missing value treatment, and aggregation logic. Selected portfolio-level insights are presented in this section, while detailed and technical analysis are provided in the repository.

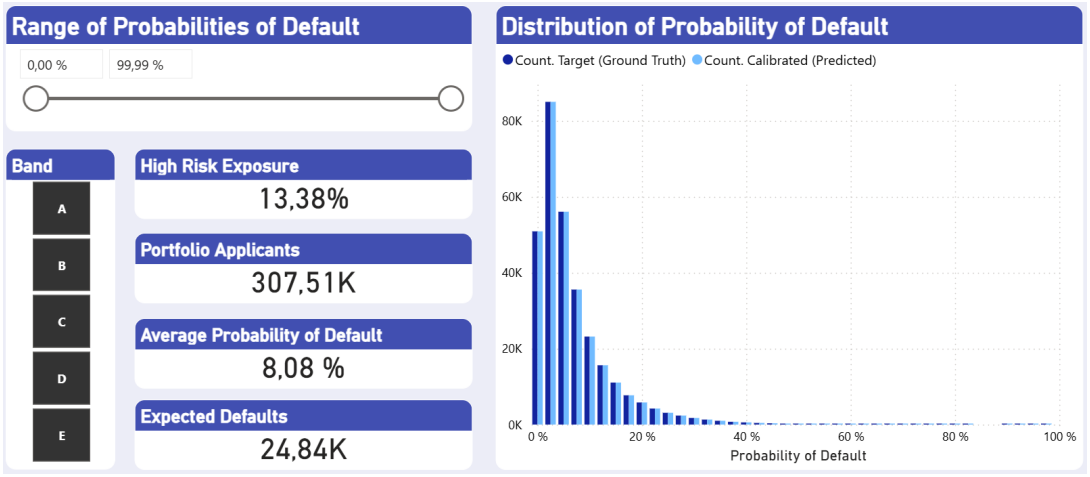

Figure 2: Portfolio overview and distribution of calibrated probabilities of default.

3.1 Portfolio Overview

The portfolio consists of approximately 307,000 loan applicants, with an observed default rate of roughly 8%, indicating a highly imbalanced classification problem. Most applicants exhibit low to moderate predicted probabilities of default, while a smaller subset concentrates significantly higher risk.

Figure 2 illustrates the distribution of PDs across the portfolio. The distribution is strongly right-skewed, with the majority of applicants clustered at low PD values and a long tail of higher-risk cases. This skewness highlights the importance of probability calibration and risk segmentation, as small changes in approval thresholds can have a disproportionate impact on portfolio risk.

3.2 Key Drivers

High-level analysis of applicant characteristics reveals intuitive relationships between observed features and default risk. cash loans exhibit higher observed default rates than revolving products (8.3% vs. 5.5%), and regions with higher observed default rates show approximately 2× the portfolio PD compared to lower-risk regions. A difference is also observed across gender groups (male applicants show roughly +3 pp higher default rate than female applicants).

Socio-economic characteristics (employment status, education, number of children, and occupation) further differentiate risk levels, consistent with expected affordability and stability effects. Bureau-derived features add strong signal: most external loans in the bureau history are closed, while higher active exposure and delinquency-related indicators are associated with higher PD. Finally, `EXT_SOURCE` variables emerge as among the most informative predictors, aligning with their interpretation as external credit indicators.

4 Modeling and Calibration

Different classification models were evaluated to estimate the Probability of Default (PD) at the applicant level. Model selection considered not only predictive performance, but also interpretability, stability, scalability, and suitability for use in regulated credit decision environments.

4.1 Candidate Models

Three model families were evaluated:

- **Logistic Regression:** a linear probabilistic model widely used in classification tasks due to its interpretability, stability, and regulatory acceptance.
- **Random Forest:** an ensemble tree-based model capable of capturing non-linear relationships and feature interactions.
- **Support Vector Machine (SVM):** a margin-based classifier implemented using stochastic gradient descent to scale to large datasets.

All models were trained and evaluated on the same curated dataset, using the area under the ROC curve (AUC) as the primary discrimination metric.

4.2 Model Comparison

All three models achieved comparable discrimination performance, with AUC values in the range of approximately 0.72–0.74. Non-linear models such as Random Forest and SVM showed marginal performance differences but did not deliver a material improvement over Logistic Regression.

Random Forest models exhibited stronger in-sample performance but showed signs of overfitting. Support Vector Machines demonstrated reasonable recall of default events but produced outputs that are not naturally probabilistic, complicating downstream interpretation and policy design.

In contrast, Logistic Regression provided stable and consistent performance across training and validation datasets while directly producing probability estimates that are well-suited for PD modeling and scorecard construction.

Table 1: Performance summary by model

Model	Train AUC	Test AUC	Recall	Best Threshold
Logistic Regression	74.27%	72.63%	67%	50.28%
Random Forest	79.13%	72.30%	66%	47.29%
SVM	74.41%	72.93%	70%	7.70%

4.3 Final Model Selection

Logistic Regression was selected as the final production model due to its strong balance between predictive performance, interpretability, and operational suitability. Although more complex models offered comparable discrimination, their additional complexity did not translate into meaningful business value for this use case.

4.4 Probability Calibration

While discrimination metrics such as AUC measure a model's ability to rank applicants by risk, they do not guarantee that predicted probabilities accurately reflect observed default rates. Miscalibrated probabilities can lead to systematic underestimation or overestimation of portfolio risk.

To address this, probability calibration techniques were evaluated, including Platt Scaling and Isotonic Regression. Both methods substantially improved probability accuracy as measured by the Brier score, while preserving discrimination performance.

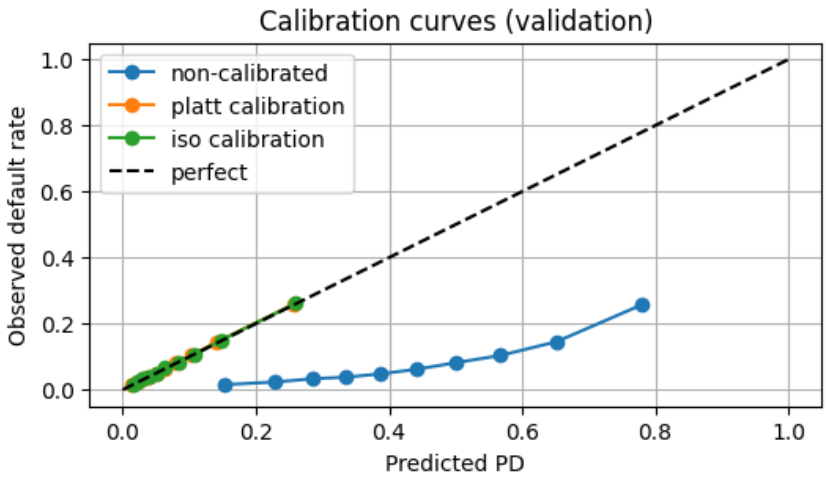

Figure 3: Calibration curves comparing non-calibrated, Platt-scaled, and isotonic-calibrated probabilities of default against observed default rates on the validation set.

Avg. PD (Grouth Truth)	Avg. PD (non-calibrated)	Avg. PD (calibrated)
8,07 %	43,43 %	8,08 %

Figure 4: Comparison of average portfolio Probability of Default (PD) across ground truth, non-calibrated model and calibrated predictions.

Figures 3 and 4 illustrate the impact of calibration at both the individual prediction level.

5 Scorecard and Risk Bands

To translate continuous Probability of Default (PD) estimates into actionable decision tools, the calibrated model outputs were transformed into a scorecard and risk bands.

5.1 Band Definition

Calibrated PDs were first transformed into a monotonic risk score using the following relationship:

$$\text{Score} = -\ln(\text{PD})$$

This transformation ensures an intuitive ordering, where lower probabilities of default correspond to higher scores and better credit quality. The score preserves the relative ranking of applicants while improving numerical stability and interpretability.

Applicants were then assigned to five risk bands (A–E). Table 2 summarizes the band structure.

Table 2: Risk band definition based on calibrated PD

Band	Calibrated PD Range
A	$[0.00, 0.01]$
B	$(0.01, 0.03]$
C	$(0.03, 0.07]$
D	$(0.07, 0.15]$
E	$(0.15, 1.00]$

5.2 Monotonicity and Stability

A key requirement for scorecard-based credit systems is monotonicity: higher risk bands must consistently exhibit higher default rates.

Table 3: Scorecard summary by risk band

Band	Avg. Calibrated PD	Avg. Grouth truth PD	Composition
A	∼0.8%	∼1.1%	4.30K
B	∼2.1%	∼2.0%	65.61K
C	∼4.8%	∼4.6%	112.97K
D	∼10.1%	∼10.4%	83.48K
E	∼23.25%	∼23.19%	41.14K

Thresholds are illustrative and can be tuned to align with business risk appetite, acceptance targets, and observed portfolio performance.

5.3 Interpretation of Risk Bands

Each risk band represents a distinct credit quality segment with clear business interpretation. Applicants in Bands A and B exhibit very low default risk and are suitable for automatic approval under most risk appetites. Band C represents moderate risk and is typically associated with conditional approval.

Bands D and E correspond to high-risk applicants, with substantially elevated default probabilities. These bands are commonly associated with tighter credit conditions, higher pricing, enhanced review processes, or outright rejection depending on institutional risk appetite.

5.4 Governance and Practical Use

The scorecard and risk band framework supports governance requirements by providing a transparent link between model outputs and business decisions because the band definitions are based on calibrated PDs and exhibit strong monotonicity, the framework is well-suited for integration into approval systems, reporting dashboards, and risk management processes.

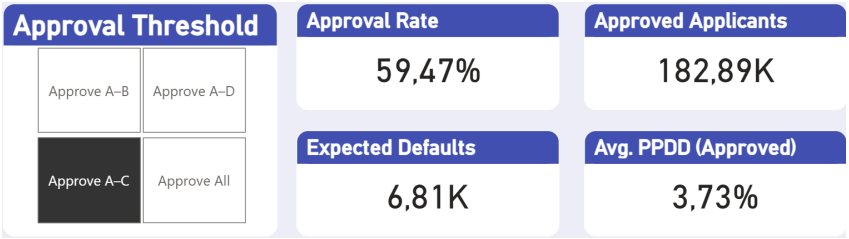

Figure 5: Approval rates and average portfolio risk across the simulated A-C strategy.

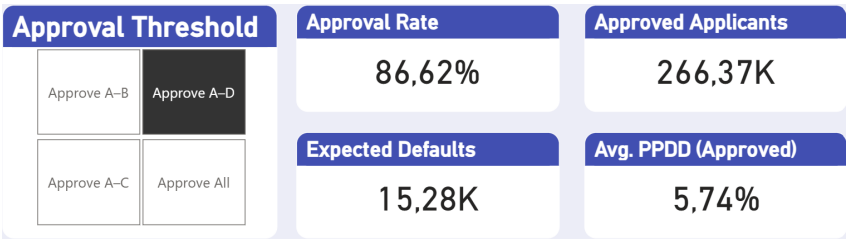

Figure 6: Approval rates and average portfolio risk across the simulated A-D strategy.

6 Credit Strategy Simulation

To demonstrate the practical value of the scorecard and risk band framework, some approval strategies were simulated based on different combinations of risk bands.

6.1 Approval Policies

Three illustrative approval strategies were evaluated:

- **Conservative Strategy (A-C):** approve applicants in Bands A, B, and C only.
- **Moderate Strategy (A-D):** approve applicants in Bands A through D.
- **Aggressive Strategy (A-E):** approve all applicants regardless of risk band.

These strategies enable a transparent comparison between portfolio growth and credit risk, using predicted PDs and observed default behavior.

Under the Conservative Strategy (A-C) (Figure 5), the approval rate is approximately 59%, while maintaining a low average Probability of Default below 4%. This strategy significantly limits exposure to high-risk applicants and concentrates approvals in lower-risk segments.

Expanding approvals to include Band D (Figure 6) increases the approval rate but leads to a disproportionate increase in expected defaults. The Aggressive Strategy, while maximizing approval volume, results in substantially higher portfolio risk.

- Moving from A–C to A–D increases approvals by +27.15pp (59.47% to 86.62%).
- Expected defaults increase by +8.47K (6.81K to 15.28K), about 2.24× higher.
- Avg. PD increases by +2.01pp (3.73% to 5.74%).

6.2 Concentration of Risk

Although higher-risk bands represent a smaller share of total applicants, they account for a disproportionately large share of risk exposure. Figure 7 illustrates how risk exposure is concentrated in Bands D and E.

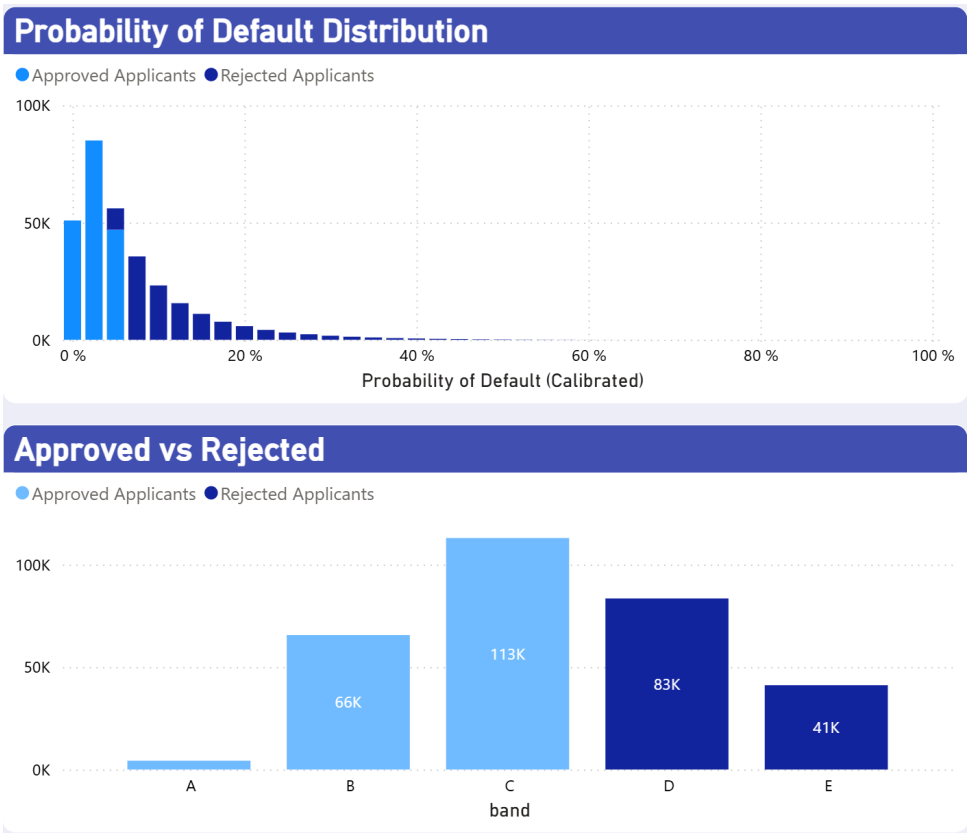

Figure 7: Concentration of risk exposure across risk bands Approved vs rejected distribution under the A–C strategy, illustrating concentration of higher PD applicants in rejected bands and higher PD ranges.

This concentration effect explains why relatively small changes in approval thresholds can materially affect portfolio performance. Excluding the highest-risk bands leads to a significant reduction in risk exposure.

6.3 Business Interpretation

The simulation results highlight the role of calibrated PDs and risk bands as decision tools for aligning portfolio growth with risk appetite. Under the conservative A–C policy, the portfolio approves approximately 59% of applicants while maintaining an average approved PD of approximately 3.7% and an expected default count of approximately 6.8K. Expanding approvals to include Band D increases approval volume to approximately 86.6%, but portfolio risk rises materially (average approved PD around 5.7%) and expected defaults increase to approximately 15.3K.

7 Portfolio Insights

The calibrated PDs and risk bands enable continuous monitoring of portfolio risk at both aggregate and segment levels. Power BI dashboards were developed to provide a consolidated view of portfolio composition, risk concentration, and model behavior, supporting ongoing oversight and decision-making.

7.1 Explainability and Risk Drivers

Model explainability usually is a requirement in credit risk applications. The selected Logistic Regression model supports interpretation of feature effects, enabling risk managers to understand the primary drivers of default risk and assess whether the model behavior aligns with domain intuition.

Across model diagnostics and portfolio segmentation views, default risk is most strongly associated with external credit score indicators (e.g., EXT_SOURCE features), measures of existing credit exposure and repayment capacity, and applicant socio-economic characteristics. Applicants with weaker bureau indicators, higher leverage, and more adverse credit histories consistently exhibit higher calibrated PDs.

7.2 Segment Monitoring and Portfolio Hotspots

Beyond individual-level explainability, segment-level monitoring supports governance and portfolio management. The dashboard enables a breakdown of calibrated PDs across key segments (e.g., occupation type, education level, age group, and credit amount), helping identify concentration areas where high-risk exposure is elevated. Furthermore, Monitoring these segment trends over time provides early warning signals for the portfolio quality.

8 Limitations

While the proposed framework demonstrates a first-step approach to PD modeling and credit strategy simulation, several limitations should be acknowledged.

- The analysis is based on a static snapshot of applicant information at the time of loan origination. The absence of an explicit time dimension prevents modeling of risk migration, cohort performance, and lifecycle behavior across multiple periods, and limits the ability to evaluate temporal stability of the score over time.
- The dataset used in this project is inspired by a real-world credit portfolio. While the structure and relationships reflect realistic credit risk patterns, results should be interpreted as illustrative rather than validated on proprietary bank performance data.
- The solution focuses on Probability of Default (PD) and does not model Loss Given Default (LGD) or Exposure at Default (EAD). As a result, the framework supports PD-based decisioning and portfolio steering but does not produce full expected-loss estimates (EL = PD $\times$ LGD $\times$ EAD).

9 Next Steps

- Expand data sources: Incorporate additional portfolio and behavioral datasets beyond applications and bureau (e.g., repayment history, prior applications, and revolving credit behavior) can be integrated into the ETL layer to enrich behavioral and repayment signals.
- Evaluate more advanced models: Benchmark gradient-boosted decision trees (e.g., LightGBM/XGBoost) and other non-linear models against Logistic Regression.
- Extend to expected loss: Add LGD and EAD components to evolve from PD-only decisioning into expected-loss and capital-aware portfolio management.

References

Amazon Web Services (2025). *AWS Documentation*. Online documentation. Accessed: 2025. URL: https://docs.aws.amazon.com.

Home Credit Group (2018). *Home Credit Default Risk*. Kaggle dataset. Accessed: 2025. URL: https://www.kaggle.com/c/home-credit-default-risk.

Scikit-learn developers (2025). *Supervised Learning*. scikit-learn documentation. Accessed: 2025. URL: https://scikit-learn.org/stable/supervised_learning.html.