Professional Certificate in AI for Credit Management

Machine Learning Techniques for Credit Scoring

Credit scoring is the quantitative process of estimating the likelihood that a borrower will fulfill their financial obligations. In the context of machine learning, a wide array of technical terms and concepts are used to build, evaluate, and maintain predictive models. This guide presents the essential vocabulary that learners of the Professional Certificate in AI for Credit Management need to master. Each term is defined, illustrated with practical examples, and linked to common challenges that arise in real-world credit environments.

Target variable – The outcome that a model is trained to predict. In credit scoring the target is typically a binary indicator such as default (1) or non-default (0). For example, a bank may define default as a loan that is 90 days past due.

Predictor – Also called a feature or independent variable, a predictor is any piece of information that may help explain the target. Common predictors include credit history length, debt-to-income ratio, and number of recent inquiries.

Feature – The term used interchangeably with predictor, but often refers to the processed version of raw data that is actually fed into a model. For instance, a raw field “annual income” may be transformed into a logarithmic feature to reduce skewness.

Label – Another name for the target variable, especially when discussing supervised learning. In a dataset of loan applications, each row is labeled as 1 for default or 0 for no default.

Observation – A single record or instance in a dataset, representing one borrower’s profile and outcome.

Dataset – A collection of observations, typically stored in a tabular format where rows are observations and columns are features and the label.

Supervised learning – A class of algorithms that learn a mapping from predictors to a target using labeled data. Credit scoring models are almost always supervised because historical repayment outcomes are known.

Unsupervised learning – Techniques that discover structure in data without explicit labels. In credit management, clustering borrowers into risk buckets can be an unsupervised task that informs segmentation strategies.

Semi-supervised learning – Methods that combine a small labeled set with a larger unlabeled set. This approach can be useful when recent loan data are available but outcomes are still pending.

Classification – The type of supervised learning where the target is categorical. Credit scoring is a binary classification problem, but multi-class classification may be used when distinguishing among several risk levels (e.g., low, medium, high).

Regression – Supervised learning where the target is continuous. In credit scoring, regression models are sometimes employed to predict the probability of default (a continuous value between 0 and 1) before applying a threshold.

Logistic regression – A statistical model that estimates the probability of a binary outcome using a logistic function. It is a baseline model for credit scoring because its coefficients can be interpreted as odds ratios. Example: a coefficient of 0.7 for “past due accounts” means each additional past-due account multiplies the odds of default by e^0.7 ≈ 2.01.

Decision tree – A flow-chart-like structure that recursively splits the feature space based on values that maximize information gain or reduce impurity. Trees capture non-linear relationships and interactions without explicit feature engineering. A simple tree might first split on “credit utilization > 80%” and then on “number of open credit lines > 5”.

Random forest – An ensemble of decision trees built on bootstrapped samples of the data, with each tree considering a random subset of features at each split. Random forests reduce variance compared with a single tree and are robust to overfitting. In credit scoring, they can model complex patterns while still providing variable importance scores.

Gradient boosting machine (GBM) – An ensemble method that builds trees sequentially, each new tree correcting the errors of the combined previous trees. GBM often yields higher predictive accuracy than random forest but requires careful tuning to avoid over-fitting.

XGBoost – An optimized implementation of gradient boosting that adds regularization, parallel processing, and tree pruning. It has become a de-facto standard for many credit scoring competitions because of its speed and accuracy.

LightGBM – A gradient boosting framework that grows trees leaf-wise rather than level-wise, leading to faster training on large datasets. It also supports categorical feature handling without one-hot encoding.

CatBoost – A gradient boosting library that excels with categorical data by using ordered boosting and target statistics. It reduces the need for extensive preprocessing of categorical fields such as “employment type”.

Support vector machine (SVM) – An algorithm that finds the hyperplane that maximally separates classes in a high-dimensional space. Kernel functions allow SVMs to capture non-linear relationships. SVMs can be effective for credit scoring when the dataset is moderate in size and the feature space is well-scaled.

k-Nearest Neighbors (k-NN) – A non-parametric method that classifies an observation based on the majority class among its k closest neighbors in the feature space. While simple, k-NN can be

computationally expensive for large credit portfolios and is sensitive to the choice of distance metric.

Naïve Bayes – A probabilistic classifier that assumes independence among features. Despite its strong independence assumption, it can be surprisingly effective for high-dimensional text-based credit applications (e.g., analyzing free-form comments).

Neural network – A family of models composed of layers of interconnected nodes (neurons) that learn hierarchical representations of data. Deep neural networks can capture intricate patterns but require large training sets and careful regularization to remain interpretable.

Embedding – A dense vector representation learned for categorical variables, often using neural networks. In credit scoring, merchant category codes or geographic identifiers can be embedded to capture similarity relationships.

Feature engineering – The process of creating, transforming, and selecting variables that improve model performance. Effective feature engineering often determines the success of a credit scoring model.

One-hot encoding – Converting a categorical variable with N distinct values into N binary columns. For a field “home ownership” with values {rent, own, mortgage}, three new columns are created, each indicating the presence of one category.

Target encoding – Replacing a categorical variable with the mean of the target variable for each category, often smoothed to avoid overfitting. For example, the average default rate for each “state” can be used as a numeric feature.

Interaction term – A feature created by multiplying two or more base features to capture joint effects. In credit scoring, “high utilization * many open lines” may be a stronger predictor of default than either factor alone.

Polynomial feature – Raising a numeric feature to a higher power (e.g., squared or cubed) to enable linear models to capture curvature. Adding a squared “debt-to-income” term can help logistic regression model a U-shaped risk curve.

Binning – Grouping a continuous variable into discrete intervals. For “age”, bins such as 18-25, 26-35, 36-45, etc., can reduce noise and make the effect of age more interpretable.

Scaling – Adjusting numeric features to a common range, such as standardizing to zero mean and unit variance. Scaling is essential for algorithms that rely on distance calculations, like SVMs or k-NN.

Imputation – Filling missing values with estimated replacements, such as the median, mean, or a model-based prediction. In credit datasets, missing “annual income” values are often imputed with the median income of the applicant’s region.

Outlier detection – Identifying observations that deviate markedly from the rest of the data. Techniques

include Z-score filtering, interquartile range (IQR) rules, or model-based methods. Outliers can distort model coefficients, especially in linear models.

Dimensionality reduction – Techniques that reduce the number of features while preserving essential information, such as Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE). In credit scoring, PCA may be used to compress highly correlated financial ratios into a few principal components.

Model evaluation – The set of metrics and methods used to assess how well a model predicts the target on unseen data. In credit scoring, evaluation focuses on both discrimination (ability to separate good and bad borrowers) and calibration (alignment of predicted probabilities with observed outcomes).

Confusion matrix – A table that summarizes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For a default prediction model, TP represents correctly identified defaults, while FP represents non-defaulters incorrectly flagged as risky.

Accuracy – The proportion of correctly classified observations: (TP + TN) / (TP + TN + FP + FN). Accuracy can be misleading in credit scoring because default events are rare (often Precision – The proportion of predicted positives that are true positives: TP / (TP + FP). High precision means few good borrowers are mistakenly labeled as high risk.

Recall – Also called sensitivity or true positive rate: TP / (TP + FN). Recall measures how many actual defaults the model captures.

F1-score – The harmonic mean of precision and recall, providing a single metric that balances the two. In credit scoring, the F1-score is useful when both false positives and false negatives carry significant costs.

Area under the ROC curve (AUC-ROC) – The probability that a randomly chosen default will receive a higher predicted risk score than a randomly chosen non-default. AUC values range from 0.5 (no discrimination) to 1.0 (perfect discrimination).

Kolmogorov-Smirnov statistic (KS) – The maximum difference between the cumulative distribution functions of the scores for the two classes. A KS above 0.4 is often considered strong discrimination in the banking industry.

Gini coefficient – Derived from the AUC, calculated as 2 × AUC – 1. A Gini of 0.6 corresponds to an AUC of 0.8. Gini is frequently reported in credit risk reports.

Lift – The ratio of the default rate in a selected score band to the overall default rate. A lift of 3 in the top 10 % of scores means that segment has three times the average default frequency, indicating effective ranking.

Calibration – The agreement between predicted probabilities and observed default frequencies. Calibration plots (e.g., reliability diagrams) show whether a model that predicts a 10 % default probability actually

experiences a 10 % default rate in that bucket.

Cross-validation – A technique for estimating model performance by repeatedly splitting the data into training and validation folds. k-fold cross-validation (commonly k = 5 or 10) provides a more stable estimate than a single holdout set.

Time-series split – A variant of cross-validation that respects temporal ordering, ensuring that training data always precede validation data. This is crucial in credit scoring because patterns evolve over time.

Holdout set – A portion of the data set aside for final model testing, never used during training or hyper-parameter tuning. A typical split might allocate 70 % for training, 15 % for validation, and 15 % for holdout.

Hyper-parameter – A configuration setting that controls model complexity but is not learned from the data. Examples include the number of trees in a random forest, learning rate in XGBoost, or regularization strength in logistic regression.

Regularization – Techniques that penalize large model coefficients to prevent overfitting. L1 regularization (lasso) can drive some coefficients to zero, performing implicit feature selection. L2 regularization (ridge) shrinks coefficients toward zero while keeping all features.

Overfitting – When a model captures noise or idiosyncrasies in the training data, leading to poor performance on new data. Indicators include a large gap between training and validation AUC.

Underfitting – When a model is too simple to capture the underlying patterns, resulting in low performance on both training and validation sets.

Bias-variance trade-off – The balance between model simplicity (high bias) and complexity (high variance). Effective credit scoring models find a sweet spot that minimizes total error.

Model deployment – The process of integrating a trained model into production systems so that it can score live loan applications. Deployment can be batch-oriented (scoring a nightly file) or real-time (API call per application).

Scoring pipeline – The sequence of steps that transforms raw application data into a risk score. Typical stages include data ingestion, preprocessing (imputation, encoding), feature generation, model inference, and output formatting.

Model monitoring – Ongoing observation of model performance metrics (e.g., AUC, KS, calibration) after deployment. Monitoring helps detect performance drift caused by changes in borrower behavior or macro-economic conditions.

Concept drift – A shift in the statistical relationship between predictors and the target over time. In credit scoring, drift may occur when a new regulatory rule changes lending standards, altering default patterns.

Feature drift – Changes in the distribution of input features, such as an increase in the average credit utilization across the portfolio. Feature drift can affect model predictions even if the underlying relationship remains stable.

Shadow testing – Running a new model in parallel with the existing production model without affecting decisions. Results are compared to evaluate potential improvements before full rollout.

Explainability – The capacity to articulate why a model gave a particular prediction. Techniques such as SHAP values, LIME, or global feature importance charts support explainability, which is essential for regulatory compliance and stakeholder trust.

Model risk management – A governance framework that includes documentation, validation, back-testing, and periodic review of credit scoring models. The framework ensures models meet internal standards and external regulations.

Fairness – The principle that a model should not produce discriminatory outcomes based on protected attributes such as race, gender, or age. Fairness metrics include demographic parity, equal opportunity, and disparate impact.

Regulatory compliance – Adherence to laws and guidelines such as the Basel III framework, the Equal Credit Opportunity Act (ECOA), and local data-privacy statutes. Compliance often mandates documentation of model development, validation, and explainability.

Class imbalance – The situation where one class (typically non-default) vastly outnumbers the other. Imbalance can cause models to be biased toward the majority class, inflating accuracy while missing defaults.

Resampling – Techniques to address class imbalance, including oversampling the minority class (e.g., SMOTE) or undersampling the majority class. Oversampling creates synthetic default cases, while undersampling discards some non-default records.

Cost-sensitive learning – Incorporating different misclassification costs directly into the training objective. For credit scoring, a false negative (missed default) may be weighted higher than a false positive (rejected good borrower).

Threshold selection – Choosing a cut-off probability that converts continuous risk scores into binary decisions (approve/reject). The optimal threshold balances business objectives such as profit, risk appetite, and regulatory limits.

Profit curve – A plot that shows expected profit as a function of the acceptance rate or score threshold. It helps identify the threshold that maximizes net profit while respecting risk limits.

Portfolio segmentation – Grouping borrowers into risk tiers based on score bands, allowing differentiated pricing, underwriting, and monitoring strategies.

Stress testing – Simulating adverse economic scenarios (e.g., recession, high unemployment) to assess how the credit portfolio would perform under extreme conditions. Models may be re-scored with stressed macro variables to estimate potential losses.

Loss given default (LGD) – The proportion of exposure that is not recovered when a borrower defaults. LGD models often use regression or classification techniques to predict recovery rates based on collateral, seniority, and macro variables.

Exposure at default (EAD) – The estimated outstanding balance at the time of default. EAD modeling may involve forecasting future draws on revolving credit lines.

Probability of default (PD) – The output of a credit scoring model, representing the likelihood that a borrower will default within a specified horizon (e.g., 12 months).

Risk-adjusted return – A performance metric that combines expected profit with the risk of default, often expressed as risk-adjusted return on capital (RAROC).

Data leakage – The inadvertent inclusion of information that would not be available at the time of prediction, leading to overly optimistic performance estimates. An example is using “days past due at month end” as a predictor for a loan that is still active.

Multicollinearity – High correlation among predictors, which can inflate variance of coefficient estimates in linear models. Techniques such as variance inflation factor (VIF) analysis or dimensionality reduction can mitigate multicollinearity.

Feature importance – A ranking that indicates how much each predictor contributes to model predictions. In tree-based models, importance can be measured by gain, split count, or permutation impact.

Permutation importance – Assessing feature importance by randomly shuffling a single feature’s values and measuring the drop in model performance. This method works for any model type and highlights the true predictive contribution of each feature.

Partial dependence plot (PDP) – A visualization that shows the marginal effect of a feature on the predicted outcome, averaged over the distribution of all other features. PDPs help interpret non-linear relationships captured by complex models.

SHAP values – A unified framework that attributes a model’s prediction to each feature based on game-theoretic concepts. SHAP provides both global (overall) and local (individual) explanations, making it a preferred tool for credit model interpretability.

Model governance – The set of policies, procedures, and controls that oversee model lifecycle activities, from data acquisition to model retirement. Governance ensures consistency, accountability, and auditability.

Example: logistic regression in practice

A mid-size bank wants to predict 12-month default for new personal loans. The data scientist assembles a dataset of 150,000 historic loans, each with the following features:

- Age (numeric)
- Annual income (numeric)
- Credit utilization (percentage)
- Number of past due accounts (integer)
- Home ownership (categorical: rent, own, mortgage)
- Loan purpose (categorical: debt-consolidation, home-improvement, other)
- Length of credit history (months)

After cleaning, missing income values are imputed with the median income of the applicant's zip code. Categorical variables are one-hot encoded, resulting in eight binary columns. The target variable is default = 1 if the loan is 90 days past due within 12 months, otherwise 0.

The analyst fits a logistic regression with L2 regularization (C = 0.5). Coefficients reveal that "credit utilization > 80 %" carries a weight of 1.2 (odds ratio ≈ 3.3), while "home ownership = own" has a weight of -0.4 (odds ratio ≈ 0.67). The model achieves an AUC of 0.78 on a holdout set, and calibration analysis shows that predicted probabilities align closely with observed default rates in decile buckets.

Because logistic regression provides transparent coefficients, the bank can document the influence of each factor, satisfying regulatory expectations for explainability.

Example: gradient boosting with target encoding

A fintech startup processes a high volume of small-ticket credit-card applications. The dataset includes a categorical variable "merchant category code" (MCC) with over 300 distinct values, many of which appear only a few times. Direct one-hot encoding would explode the feature space and lead to sparse data.

The data science team applies target encoding to MCC, computing the smoothed default rate for each code. To avoid leakage, the encoding is calculated using only the training folds within a cross-validation loop. After encoding, they train a LightGBM model with the following hyper-parameters:

- num_leaves = 31
- learning_rate = 0.05
- max_depth = -1 (no explicit depth limit)
- feature_fraction = 0.8

Cross-validation yields an AUC of 0.84, a KS of 0.42, and a Gini of 0.68. Feature importance shows that the target-encoded MCC, credit utilization, and recent inquiry count are the top three contributors.

The model is deployed as a real-time scoring API. Each incoming application triggers the same

preprocessing pipeline: missing values are imputed, MCC is target-encoded using the pre-computed statistics, and the LightGBM model returns a default probability. The fintech monitors the live AUC weekly; after a three-month period, a slight decline (AUC = 0.81) prompts an investigation that uncovers a change in the macro-economic environment, leading to a scheduled model retraining.

Practical challenges in credit-scoring machine learning

1. Data quality and completeness – Credit datasets often contain missing fields, inconsistent formats, and legacy codes. Robust preprocessing pipelines that handle imputation, standardization, and validation are essential.

2. Class imbalance – Default events are rare, causing models to be biased toward the majority class. Techniques such as SMOTE, cost-sensitive loss functions, and careful threshold selection are required to ensure the model captures the minority class effectively.

3. Regulatory constraints on features – Certain variables (e.g., race, gender) are prohibited for credit decisioning. Feature selection must respect legal restrictions while still capturing predictive power.

4. Explainability vs. performance trade-off – Highly accurate models like deep neural networks may be opaque. Organizations must balance the desire for predictive accuracy with the need for transparent explanations that regulators and customers can understand.

5. Concept drift and model decay – Economic cycles, policy changes, or shifts in consumer behavior can erode model performance over time. Continuous monitoring, periodic retraining, and automated alerts for drift are necessary to maintain model relevance.

6. Integration with legacy systems – Credit scoring models often need to interface with mainframe-based underwriting platforms. Ensuring seamless data flow, latency constraints, and security compliance can be technically demanding.

7. Computational resources – Training large ensembles (e.g., XGBoost with thousands of trees) on millions of records may require distributed computing or cloud resources. Efficient hyper-parameter search strategies (e.g., Bayesian optimization) help manage compute costs.

8. Fairness and bias mitigation – Even when protected attributes are excluded, proxy variables can still produce disparate impact. Bias detection tools, fairness constraints during training, and post-processing adjustments (e.g., equalized odds) are part of a responsible AI toolkit.

9. Model governance documentation – Detailed records of data sources, preprocessing steps, model architecture, validation results, and versioning are required for audits. Automated documentation pipelines can reduce manual effort and improve traceability.

10. Interpretation of complex feature interactions – Tree ensembles capture high-order interactions that are difficult to articulate. Techniques such as SHAP interaction values or rule-extraction algorithms help

translate these patterns into business-friendly insights.

Key vocabularies related to model validation and performance tracking

Back-testing – Evaluating a model using historical data that was not part of the training set, often by simulating how the model would have performed if deployed at a past date.

Out-of-time validation – A validation approach where the test set consists of data from a later time period than the training data, mimicking real-world deployment conditions.

Bootstrap sampling – Drawing random samples with replacement from the original dataset to create multiple training sets, used for estimating the variability of model performance.

Confidence interval – A range that quantifies the uncertainty around a performance metric, such as AUC ± 0.02 at a 95 % confidence level.

Statistical significance testing – Methods such as DeLong's test for comparing AUCs of two models to determine whether observed differences are likely due to chance.

Model versioning – Assigning unique identifiers to each iteration of a model, along with metadata describing training data, hyper-parameters, and performance metrics.

Model registry – A centralized repository that stores model artifacts, version information, and lineage, facilitating reproducibility and governance.

Production monitoring dashboard – A visual interface that tracks key indicators (e.g., daily default rate, score distribution, latency) to alert stakeholders of anomalies.

Drift detection metric – Quantitative measures such as Population Stability Index (PSI) or Jensen-Shannon divergence that compare the distribution of a feature or score between training and current data.

Alert threshold – Pre-defined limits for drift metrics that trigger investigations or automated model retraining when exceeded.

Advanced topics and emerging techniques

AutoML – Automated machine learning platforms that handle feature preprocessing, model selection, and hyper-parameter tuning with minimal human intervention. AutoML can accelerate model development but still requires expert oversight for fairness and interpretability.

Transfer learning – Leveraging a model trained on a related task (e.g., fraud detection) as a starting point for credit scoring, especially useful when the target dataset is small.

Ensemble stacking – Combining predictions from multiple base learners (e.g., logistic regression, random

forest, XGBoost) using a meta-learner to improve overall accuracy.

Survival analysis – Modeling time-to-event data, such as the hazard of default over time, using techniques like Cox proportional hazards or deep survival models. This approach provides richer information than a static binary classification.

Explainable AI (XAI) frameworks – Libraries such as SHAP, LIME, and IBM AI Explainability 360 that standardize the generation of model explanations, supporting compliance with emerging AI regulations.

Federated learning – Training models across multiple institutions without sharing raw data, preserving privacy while leveraging a broader data pool. In credit scoring, banks could collaborate to improve risk models without exposing proprietary customer information.

Graph-based credit networks – Representing borrowers and their relationships (e.g., co-signers, shared addresses) as graphs, then applying Graph Neural Networks (GNNs) to capture network effects on credit risk.

Explainable boosting machine (EBM) – An interpretable model that builds additive, shape-function terms for each feature, offering performance close to black-box models while remaining transparent.

Model compression – Techniques such as pruning, quantization, or knowledge distillation that reduce model size and inference latency, important for real-time scoring on low-power devices.

Privacy-preserving techniques – Methods like differential privacy or homomorphic encryption that enable scoring on encrypted data, aligning with stringent data-protection regulations.

Glossary of frequently encountered terms

Binary classification – Predicting one of two possible outcomes.

Multiclass classification – Predicting one of three or more categories.

Feature matrix – The two-dimensional array (often denoted X) containing all predictor values for the observations.

Label vector – The one-dimensional array (often denoted y) containing the target values.

Training set – The subset of data used to fit model parameters.

Validation set – The subset used to tune hyper-parameters and prevent overfitting.

Test set – The final holdout data used to assess generalization performance.

Hyper-parameter tuning – Systematic search (grid, random, Bayesian) for optimal model settings.

Grid search – Exhaustive enumeration of hyper-parameter combinations.

Random search – Sampling a fixed number of hyper-parameter configurations from a predefined distribution.

Bayesian optimization – A probabilistic approach that models the performance surface and selects hyper-parameters to explore promising regions.

Early stopping – Halting training when validation performance ceases to improve, preventing overfitting in iterative algorithms such as gradient boosting.

Learning rate – A scalar that controls the step size during model updates; smaller values lead to slower but more stable learning.

Regularization strength – The magnitude of the penalty applied to model coefficients; higher values increase shrinkage.

Feature importance score – A numeric value indicating the relative contribution of a feature to the predictive power of the model.

Permutation test – A statistical test that shuffles labels to assess whether a model's performance is better than chance.

Model drift – The degradation of model performance over time due to changes in data or environment.

Population Stability Index (PSI) – A metric that compares the distribution of a variable between two samples; values above 0.25 typically signal significant drift.

Jensen-Shannon divergence – A symmetric measure of divergence between probability distributions, often used for drift detection.

Ensemble – A combination of multiple models whose predictions are aggregated (e.g., by averaging or voting) to improve robustness.

Bagging – Bootstrap aggregating, a technique that builds multiple models on different bootstrap samples and averages their predictions.

Boosting – Sequentially training models, each focusing on the errors of its predecessor, to reduce bias.

Stacking – A meta-learning technique that learns how to best combine predictions from diverse base models.

Calibration curve – A plot that compares predicted probabilities with observed frequencies, used to assess probability accuracy.

Reliability diagram – Another term for a calibration curve, often displaying error bars for each probability bin.

Disparate impact – A fairness metric that measures the ratio of favorable outcomes between protected and

unprotected groups; a ratio below 0.8 may indicate potential discrimination.

Equalized odds – A fairness condition requiring that true positive and false positive rates be equal across groups.

Demographic parity – A fairness condition requiring that the selection rate be the same across groups, regardless of underlying risk differences.

Feature selection – The process of choosing a subset of relevant predictors, using methods such as recursive feature elimination, L1 regularization, or mutual information.

Recursive feature elimination (RFE) – An iterative approach that removes the least important features based on model performance until a desired number of features remains.

Mutual information – A statistical measure of the dependence between two variables, useful for selecting features that share information with the target.

Variance Inflation Factor (VIF) – A diagnostic that quantifies multicollinearity; VIF values above 10 often warrant investigation.

Outlier – An observation that deviates markedly from the rest of the data, potentially influencing model estimates.

Robust scaling – Scaling based on median and interquartile range, less sensitive to outliers than standard scaling.

Time-dependent validation – Validation that respects chronological order, essential for credit scoring where future data should not inform past predictions.

Model interpretability – The degree to which a human can understand the internal mechanics of a model and its predictions.

Explainable AI (XAI) – A set of techniques that provide insights into model decisions, often required for regulatory compliance.

Model audit – A systematic review of a model's development, performance, and governance to ensure compliance and reliability.

Model retirement – The process of decommissioning an outdated model, archiving its artifacts, and replacing it with a newer version.

Illustrative workflow for building a credit-scoring model

1. Data acquisition – Pull loan application data, credit bureau reports, and macro-economic indicators from the data warehouse.

2. Exploratory data analysis – Summarize distributions, detect missingness patterns, and compute correlation matrices.

3. Preprocessing – Impute missing values, encode categorical variables (one-hot for low-cardinality fields, target encoding for high-cardinality fields), and scale numeric features.

4. Feature engineering – Create interaction terms (e.g., utilization × number of open lines), bin age groups, and generate rolling credit-history metrics (e.g., average balance over the past 6 months).

5. Train-validation split – Use an out-of-time split: training on loans originated before 2023-01-01, validation on loans from 2023-01-01 to 2023-06-30, and holdout on loans after 2023-07-01.

6. Model selection – Fit baseline logistic regression, random