Certificate in Credit Risk Analytics in Python

Introduction To Machine Learning

Algorithm – A step-by-step procedure for solving a problem or performing a computation. Related terms: model, training, optimization. Explanation: In machine learning, an algorithm defines how a model learns from data, adjusts its parameters, and makes predictions. For credit risk, a common algorithm is logistic regression, which estimates the probability of default based on borrower attributes. Example: Using the gradient descent algorithm to minimize the loss function of a neural network that predicts loan delinquency. Challenges: Choosing an algorithm that balances predictive accuracy with interpretability, especially when regulators require transparent risk models.

Artificial Neural Network (ANN) – A computational model inspired by the structure of biological neurons. Related terms: deep learning, layers, activation function. Explanation: ANNs consist of interconnected nodes (neurons) organized in layers; each connection carries a weight that is learned during training. In credit risk analytics, ANNs can capture complex, non-linear relationships among financial variables. Example: A feed-forward network with three hidden layers used to score credit card applications. Challenges: Overfitting to historical data, long training times, and difficulty in interpreting the learned representations for compliance reporting.

Bias-Variance Trade-off – The balance between a model's error due to erroneous assumptions (bias) and error due to sensitivity to fluctuations in the training set (variance). Related terms: underfitting, overfitting, generalization. Explanation: High bias leads to systematic errors and underfitting, while high variance causes the model to chase noise and overfit. In credit risk, a model that is too simple may miss important risk drivers; a model that is too complex may perform poorly on new loan portfolios. Example: Comparing a linear discriminant analysis (low variance, high bias) with a random forest (low bias, higher variance). Challenges: Selecting the right level of model complexity and applying techniques such as cross-validation, regularization, or ensemble methods to achieve a suitable trade-off.

Classification – A type of supervised learning where the output variable is categorical. Related terms: binary classification, multiclass classification, confusion matrix. Explanation: In credit risk, classification models predict discrete outcomes such as "default" vs. "Non-default". The model learns a decision boundary that separates classes based on input features. Example: Using a support vector machine to classify borrowers into "high risk" and "low risk" groups. Challenges: Imbalanced class distributions (defaults are rare), selecting appropriate evaluation metrics (e.G., AUC-ROC), and ensuring that the classification threshold aligns with business risk appetite.

Cross-Validation – A technique for assessing a model's ability to generalize to unseen data by partitioning the dataset into training and validation subsets multiple times. Related terms: k-fold, hold-out, performance metric. Explanation: In k-fold cross-validation, the data are divided into k equal parts; each part is used once as a validation set while the remaining k-1 parts form the training set. This provides a robust estimate of out-of-sample performance, essential for credit risk model validation. Example: Conducting 5-fold

cross-validation on a dataset of mortgage applications to evaluate a gradient-boosted tree model. Challenges: Computational cost for large datasets, leakage risk if preprocessing steps are not confined to the training folds, and maintaining temporal consistency when data have a time component.

Decision Tree – A flow-chart-like structure where internal nodes represent tests on features, branches represent outcomes, and leaf nodes represent predictions. Related terms: entropy, gini impurity, pruning. Explanation: Decision trees partition the feature space into rectangular regions, making them easy to interpret. In credit risk, they can illustrate how variables like debt-to-income ratio and credit history drive default risk. Example: A CART (Classification and Regression Tree) that splits first on “past due amount > $500” and then on “credit score Ensemble Methods – Techniques that combine multiple base learners to produce a stronger overall predictor. Related terms: bagging, boosting, stacking. Explanation: Ensembles reduce variance (bagging) or bias (boosting) by aggregating predictions from diverse models. In credit risk, ensembles often achieve higher predictive accuracy than single models while preserving interpretability through model-level feature importance. Example: A XGBoost model that aggregates hundreds of shallow decision trees to predict loan default probability. Challenges: Increased computational demand, difficulty in interpreting the combined model, and the risk of over-optimistic performance if cross-validation is not properly applied.

Feature Engineering – The process of creating, transforming, or selecting variables that improve model performance. Related terms: feature scaling, dummy variables, dimensionality reduction. Explanation: Effective feature engineering captures domain knowledge, such as converting raw transaction timestamps into “average monthly spend”. In credit risk, engineered features often include credit utilization ratios, payment delinquency counts, or macro-economic indicators. Example: Deriving a “new credit line count” from raw credit bureau data to capture recent borrowing behavior. Challenges: Time-consuming manual work, risk of data leakage if future information is inadvertently used, and the need to maintain consistent feature pipelines across training and production environments.

Feature Scaling – Normalizing or standardizing numeric features so that they have comparable ranges. Related terms: min-max scaling, z-score normalization, regularization. Explanation: Algorithms such as k-nearest neighbors, support vector machines, and neural networks are sensitive to the magnitude of input features; scaling ensures that no single variable dominates the learning process. Example: Applying min-max scaling to the “annual income” variable so it lies between 0 and 1 before feeding it to a logistic regression model. Challenges: Selecting the appropriate scaling method, handling outliers, and ensuring that scaling parameters are computed only on training data to avoid leakage.

Gradient Descent – An iterative optimization algorithm that updates model parameters in the direction of the steepest decrease of the loss function. Related terms: learning rate, stochastic gradient descent, convergence. Explanation: In each iteration, the gradient of the loss with respect to the parameters is computed, and parameters are moved a small step (learning rate) opposite to the gradient. This process continues until convergence criteria are met. Example: Training a logistic regression model for credit default prediction using batch gradient descent. Challenges: Choosing an appropriate learning rate to avoid divergence or slow convergence, dealing with local minima in non-convex loss surfaces, and scaling to large datasets with stochastic or mini-batch variants.

Hyperparameter – A configuration setting that governs the behavior of a learning algorithm but is not learned from the data. Related terms: grid search, random search, Bayesian optimization. Explanation: Hyperparameters include the number of trees in a random forest, the depth of a decision tree, or the regularization strength in a logistic model. Proper tuning can significantly impact predictive performance in credit risk models. Example: Using a grid search to find the optimal max_depth and min_samples_leaf for a gradient-boosted classifier. Challenges: High computational cost for exhaustive searches, risk of over-fitting to validation data, and the need for automated tools to manage the search space efficiently.

Imbalanced Data – A situation where the classes of interest have very different frequencies, often seen in credit risk where defaults are rare. Related terms: SMOTE, cost-sensitive learning, precision-recall curve. Explanation: Standard accuracy metrics become misleading; models may achieve high accuracy by simply predicting the majority class. Techniques such as resampling, synthetic minority oversampling, or adjusting class weights help mitigate imbalance. Example: Applying SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic default cases before training a random forest. Challenges: Synthetic samples may not reflect real-world default patterns, altered class distributions can affect probability calibration, and regulatory expectations often require justification of any resampling strategy.

K-Nearest Neighbors (KNN) – A non-parametric algorithm that classifies a new observation based on the majority class among its k closest training instances. Related terms: distance metric, curse of dimensionality, lazy learning. Explanation: KNN stores the entire training set and performs classification at prediction time, making it simple but computationally intensive for large credit datasets. Feature scaling is essential because distance calculations are sensitive to variable magnitude. Example: Using KNN with k=5 and Euclidean distance to classify loan applicants as “default” or “non-default”. Challenges: High memory and prediction cost, degraded performance in high-dimensional spaces, and sensitivity to noisy or irrelevant features.

Logistic Regression – A linear model that predicts the probability of a binary outcome using the logistic (sigmoid) function. Related terms: odds ratio, regularization, maximum likelihood estimation. Explanation: Logistic regression estimates the log-odds of default as a linear combination of input features. It is widely used in credit risk due to its interpretability and ease of implementation. Example: Modeling the probability of default as a function of credit score, loan-to-value ratio, and employment length. Challenges: Assumes linear relationships, may underperform when interactions and non-linear effects are strong, and requires careful handling of multicollinearity among predictors.

Macro-Economic Variables – External indicators such as unemployment rate, GDP growth, or interest rates that influence credit risk at a portfolio level. Related terms: stress testing, scenario analysis, exogenous factors. Explanation: Incorporating macro-economic variables helps capture systematic risk that is not explained by borrower-specific data alone. They are essential for regulatory stress-testing frameworks. Example: Adding the national unemployment rate as a predictor in a model that scores commercial loan applications. Challenges: Lagged effects, data availability at appropriate frequencies, and potential over-fitting to historical macro trends that may not repeat.

Model Validation – The process of assessing whether a model meets performance, stability, and regulatory requirements before deployment. Related terms: back-testing, benchmarking, model risk. Explanation: Validation includes statistical tests (e.G., KS statistic), out-of-sample performance checks, and

documentation of assumptions. In credit risk, validation ensures that PD (probability of default) models are reliable for capital allocation. Example: Conducting a 12-month back-test of a PD model and comparing predicted versus observed default rates across rating buckets. Challenges: Ensuring independence of validation team, handling data drift over time, and meeting stringent regulator expectations for model governance.

Neural Network Architecture – The design of layers, nodes, and connections that defines how information flows through a neural network. Related terms: feed-forward, recurrent, convolutional. Explanation: Different architectures are suited to different data types; for tabular credit data, a simple feed-forward multilayer perceptron often suffices, while sequential transaction data may benefit from recurrent structures. Example: A three-layer perceptron with 64, 32, and 16 neurons respectively, using ReLU activations for credit scoring. Challenges: Selecting the appropriate depth and width to avoid overfitting, managing training time, and providing sufficient interpretability for audit purposes.

Out-of-Sample Performance – The ability of a model to predict accurately on data that were not used during training. Related terms: hold-out set, generalization error, predictive power. Explanation: Out-of-sample metrics such as AUC, Gini, or Brier score provide realistic expectations of how a model will behave on future loan applications. They are essential for risk management and regulatory reporting. Example: Reporting an AUC of 0.78 On a 20 % hold-out sample of recent credit card applications. Challenges: Data leakage, temporal shifts that make historical hold-out sets unrepresentative, and the need for continuous monitoring as the portfolio evolves.

Overfitting – When a model captures noise or idiosyncrasies in the training data, leading to poor performance on new data. Related terms: regularization, early stopping, cross-validation. Explanation: Overfitted models have low training error but high validation error. In credit risk, overfitting may produce overly optimistic PD estimates that fail under stress scenarios. Example: A deep neural network that achieves 99 % accuracy on training data but only 60 % on a validation set. Challenges: Detecting overfitting early, applying appropriate regularization techniques, and balancing model complexity with interpretability.

Precision – The proportion of predicted positive cases that are actually positive. Related terms: recall, F1-score, confusion matrix. Explanation: In credit risk classification, precision measures how many borrowers flagged as high-risk truly default. High precision reduces false alarms but may miss some risky borrowers. Example: A model that predicts 200 defaults, of which 150 actually default, yields a precision of 75 %. Challenges: Trade-off with recall, especially when defaults are rare; selecting thresholds that align with business risk tolerance.

Principal Component Analysis (PCA) – A dimensionality-reduction technique that transforms correlated variables into a set of uncorrelated components ordered by explained variance. Related terms: eigenvectors, variance explained, feature reduction. Explanation: PCA can compress high-dimensional credit data (e.G., Thousands of transaction features) into a smaller set of components while preserving most of the information, which speeds up model training. Example: Reducing a 100-dimensional feature set to 10 principal components that capture 95 % of total variance before feeding them to a logistic regression. Challenges: Loss of interpretability because components are linear combinations of original features, and the technique assumes linear relationships.

Probabilistic Forecast – A prediction that provides a probability distribution rather than a single point estimate. Related terms: calibration, prediction interval, risk score. Explanation: In credit risk, probabilistic forecasts enable the calculation of expected loss and capital requirements. Models such as logistic regression naturally output probabilities, but they must be calibrated to reflect true default frequencies. Example: A PD model that assigns a 2.3 % Probability of default to a borrower, which is then used in expected loss calculations. Challenges: Ensuring calibration across different score bands, handling probability clipping for extreme values, and communicating probabilistic outputs to non-technical stakeholders.

Random Forest – An ensemble of decision trees built on bootstrap samples and random subsets of features, with predictions aggregated by majority vote (classification) or averaging (regression). Related terms: bagging, feature importance, out-of-bag error. Explanation: Random forests improve stability and accuracy over single trees while retaining some interpretability through variable importance metrics. They are widely used for credit scoring because they handle non-linear interactions and missing values gracefully. Example: A random forest with 500 trees that predicts default probability for small-business loans. Challenges: Large memory footprint, reduced transparency compared with a single decision tree, and the need to tune hyperparameters such as max_features and min_samples_leaf.

Recall – The proportion of actual positive cases that are correctly identified by the model. Related terms: sensitivity, true positive rate, miss rate. Explanation: In credit risk, recall measures how many true defaults are captured by the model's high-risk predictions. High recall reduces missed defaults but may increase false positives. Example: Out of 500 true defaults, the model correctly flags 400, yielding a recall of 80 %. Challenges: Balancing recall against precision, especially when the cost of false positives (e.G., Rejecting creditworthy applicants) is high.

Regularization – Techniques that add a penalty term to the loss function to discourage overly complex models. Related terms: L1 (Lasso), L2 (Ridge), elastic net. Explanation: Regularization shrinks coefficient values, promoting simpler models that generalize better. In credit risk, L1 regularization can also perform feature selection, highlighting the most predictive variables. Example: Adding an L2 penalty to a logistic regression to prevent coefficient explosion when many correlated financial ratios are used. Challenges: Choosing the correct regularization strength (lambda), dealing with bias introduced by penalty, and ensuring that important risk drivers are not overly penalized.

ROC Curve (Receiver Operating Characteristic) – A plot of true positive rate (recall) against false positive rate at various classification thresholds. Related terms: AUC, threshold selection, diagnostic ability. Explanation: The ROC curve visualizes the trade-off between detecting defaults and incorrectly labeling good borrowers as risky. The area under the curve (AUC) summarizes overall discriminative power. Example: An AUC of 0.82 Indicates that a randomly chosen defaulted borrower will have a higher predicted risk score than a randomly chosen non-defaulted borrower 82 % of the time. Challenges: ROC curves can be misleading with highly imbalanced data; precision-recall curves may be more informative for rare default events.

Scikit-Learn – An open-source Python library that provides simple and efficient tools for data mining and machine learning. Related terms: pipeline, estimator, cross-validation. Explanation: Scikit-Learn offers implementations of many algorithms discussed in this glossary (logistic regression, random forest, SVM,

etc.) Along with utilities for preprocessing, model selection, and evaluation, making it a core toolkit for credit risk analytics in Python. Example: Using `sklearn.ensemble.RandomForestClassifier` to train a PD model and `sklearn.metrics.roc_auc_score` to compute its AUC. Challenges: Managing version compatibility, integrating with larger data pipelines (e.g., Spark), and extending beyond the library's built-in algorithms for custom credit-risk models.

Shapley Values – A game-theoretic approach to attribute the contribution of each feature to a model's prediction. Related terms: interpretability, feature importance, model agnostic. Explanation: Shapley values provide a fair distribution of the prediction among features, helping explain why a particular borrower received a high default probability. They are especially useful for complex models like gradient-boosted trees. Example: Computing SHAP values for a loan applicant and discovering that "high credit utilization" contributed +0.12 To the default probability. Challenges: Computationally intensive for large datasets, potential misinterpretation if the underlying model is poorly calibrated, and the need to translate numerical contributions into business narratives.

Sensitivity Analysis – The study of how changes in input variables affect model outputs. Related terms: what-if scenarios, elasticity, stress testing. Explanation: In credit risk, sensitivity analysis helps assess the robustness of PD estimates to shifts in macro-economic variables or borrower characteristics, informing capital allocation and policy decisions. Example: Varying the unemployment rate from 5% to 7% and observing the impact on predicted default rates across loan segments. Challenges: Selecting realistic ranges, handling interactions among variables, and presenting results in a concise manner for senior management.

Support Vector Machine (SVM) – A classification algorithm that finds the hyperplane maximizing the margin between classes, optionally using kernel functions to handle non-linear separations. Related terms: kernel trick, soft margin, support vectors. Explanation: SVMs are effective for high-dimensional credit data and can produce sparse models (few support vectors). However, they are less interpretable than logistic regression and can be sensitive to parameter choices. Example: Training an SVM with a radial basis function kernel to separate defaulters from non-defaulters in a consumer loan dataset. Challenges: Choosing the right kernel and regularization parameter C, scaling to large datasets (computationally expensive), and limited probability output without additional calibration.

Temporal Validation – A validation approach that respects the chronological order of data, training on an earlier period and testing on a later period. Related terms: time-series split, data leakage, drift detection. Explanation: Temporal validation mimics real-world deployment where models are built on historical data and applied to future borrowers. It helps uncover performance degradation due to changing economic conditions or borrower behavior. Example: Training a PD model on loan data from 2015-2018 and validating on 2019-2020 applications. Challenges: Ensuring sufficient data in each time slice, handling concept drift, and updating models regularly to maintain predictive power.

Training Set – The subset of data used to fit a machine learning model's parameters. Related terms: validation set, test set, overfitting. Explanation: The training set provides the information the algorithm needs to learn patterns; its size and representativeness directly affect model quality. In credit risk, the training set often consists of historical loan performance records. Example: Using 70% of a mortgage portfolio's historical records as the training set for a default prediction model. Challenges: Maintaining

temporal consistency (no future data in training), handling class imbalance, and ensuring that the training data reflect the current underwriting policy.

Underfitting – When a model is too simple to capture the underlying structure of the data, resulting in high bias and poor performance on both training and unseen data. Related terms: bias, model capacity, feature selection. Explanation: Underfitted credit risk models may miss key risk drivers, leading to inaccurate PD estimates. Simpler models like naïve Bayes often underfit complex financial datasets. Example: A linear model that ignores interaction effects between debt-to-income ratio and credit history, yielding low AUC. Challenges: Detecting underfitting early, increasing model complexity without causing overfitting, and ensuring interpretability remains acceptable.

Validation Set – A portion of data set aside during model development to tune hyperparameters and assess performance before final testing. Related terms: cross-validation, early stopping, model selection. Explanation: The validation set provides an unbiased estimate of how changes to the model affect out-of-sample performance, crucial for preventing over-optimistic results in credit risk modeling. Example: Holding out 15 % of the dataset as a validation set while training a gradient-boosted classifier. Challenges: Data leakage if preprocessing is applied before splitting, reduced training data size, and the need for multiple validation splits when data are limited.

Variable Importance – A measure indicating how much each feature contributes to the predictive power of a model. Related terms: feature importance, gain, permutation importance. Explanation: In tree-based models, importance can be derived from split-gain or frequency; in linear models, absolute coefficient values serve a similar purpose. Understanding importance helps regulators assess model transparency. Example: Ranking “credit score”, “payment history”, and “loan amount” as the top three contributors in a random forest PD model. Challenges: Importance can be biased toward variables with many categories or high cardinality, and importance alone does not explain the direction of influence.

Variance Inflation Factor (VIF) – A diagnostic metric that quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. Related terms: multicollinearity, correlation matrix, feature selection. Explanation: High VIF values (>10) suggest that a predictor is linearly dependent on other predictors, potentially destabilizing coefficient estimates in logistic regression. Example: Calculating VIF for “total debt” and finding a value of 12, indicating the need to drop or combine it with “debt-to-income ratio”. Challenges: Detecting hidden multicollinearity, deciding whether to remove or transform variables, and maintaining model interpretability after adjustments.

Weighted Loss Function – A loss function that assigns different weights to observations, often used to address class imbalance. Related terms: cost-sensitive learning, class weight, imbalanced data. Explanation: By penalizing misclassification of the minority class more heavily, the model becomes more attentive to defaults. Many algorithms in scikit-learn accept a `class_weight` parameter for this purpose. Example: Setting `class_weight='balanced'` in a logistic regression to automatically weight defaults inversely proportional to their frequency. Challenges: Determining appropriate weight values, preventing over-compensation that harms overall accuracy, and ensuring that weighted training does not distort probability calibration.

Y-Intercept – The constant term in a linear model that represents the predicted value when all predictors are zero. Related terms: bias term, intercept, baseline probability. Explanation: In a logistic regression for credit risk, the y-intercept captures the base log-odds of default before considering borrower-specific characteristics. Example: A logistic regression with an intercept of –4.2 Corresponds to a baseline default probability of about 1.5 % When all features are at their reference levels. Challenges: Interpreting the intercept meaningfully when features are not centered, and ensuring that the intercept does not dominate the prediction after scaling.

Z-Score Normalization – A scaling method that transforms a variable to have zero mean and unit variance. Related terms: standardization, feature scaling, outlier sensitivity. Explanation: Z-score normalization is often preferred for algorithms that assume normally distributed inputs, such as linear discriminant analysis. In credit risk, it helps align variables like “annual income” and “age” onto comparable scales. Example: Converting “age” from years to a z-score before feeding it into a support vector machine classifier. Challenges: Sensitivity to extreme outliers, the need to compute mean and standard deviation on the training set only, and potential loss of interpretability for business stakeholders.