AI-Driven Non-Financial Risk Prediction and Control Effectiveness System for Banking

Capstone Project – Post Graduate Diploma in Artificial Intelligence & Machine Learning (Dec 2024)

Submitted by: GAT

School: Asian Institute of Management (AIM) – Emeritus

Domain: Finance – Non-Financial Risk Management

This project aimed to develop an end-to-end machine learning solution for enhancing non-financial risk (NFR) management in a banking context. The primary objectives were to predict the likelihood and severity of operational risk incidents using explainable AI, and to enable proactive risk mitigation and control prioritization.

Table of Contents

1. Problem Understanding and Framing

Business and Data Science Problem

Business Context and Problem (Non-Financial Risk Management)

Banks face thousands of non-financial risk (NFR) events yearly such as operational failures, process breakdowns, system outages, fraud by employees, compliance breaches, vendor failures, cyber incidents, and human error. These events often cause:

- financial losses
- customer impact
- regulatory scrutiny and sanctions
- reputational damage

Risk managers rely heavily on backward-looking tools and lagging indicators such as Risk and Control Self-Assessments (RCSA), Key Risk Indicators (KRIs), and periodic audit reviews. While these tools are essential for governance, they provide limited capability to anticipate emerging risks before incidents materialize.

This project addresses this limitation by framing non-financial risk as a predictive analytics problem, where machine learning models are used to generate forward-looking risk signals. The goal is not merely statistical prediction, but to support decision-making for non-financial risk managers, enabling earlier intervention and more effective allocation of limited risk mitigation resources.

Data Science Problem and Task Definition

From a data science perspective, the problem is formulated as a supervised learning task with the following objectives:

1. Predicts whether a non-financial risk event will occur within the next 30 days (Event Occurrence). Classification model will be used.
    - Possible values:
        - 1 → An operational risk incident will occur in the next 30 days
        - 0 → No incident will occur in the next 30 days
    - Why Classification is appropriate here:
        - Operational incidents are events (they either occur or they do not).
        - Risk managers need a yes or no risk signal to decide which units to monitor and where to allocate controls or audits.
        - It enables risk ranking and early warning systems.

2. Estimates the potential severity of such events (Severity Amount). Regression model will be used.

    - Possible values: any non-negative number representing financial loss amount
    - Why Regression is appropriate here:
        - Operational risk losses vary widely in magnitude.
        - Risk managers must prioritize high-impact risks, allocate remediation budgets, support capital and risk appetite discussions.
        - Regression supports impact-based decision-making.

3. Explains model decisions in a transparent and auditable manner.

Success Metrics

Technical KPIs

To evaluate the performance of the proposed AI-driven non-financial risk framework, a set of technical KPIs was defined for both the classification and regression tasks. These metrics were selected to reflect the characteristics of rare-event risk data and to ensure alignment with operational risk decision-making.

Classification KPIs: (Event Occurrence within 30 days)

- ROC-AUC (Receiver Operating Characteristic – Area Under Curve)
  Measures the model's overall ability to discriminate between event and non-event cases across all thresholds. While informative at a high level, its interpretability is limited in highly imbalanced datasets.
- PR-AUC (Precision–Recall Area Under Curve)
  Provides a more appropriate assessment of performance under class imbalance by focusing on the trade-off between precision and recall for the minority (event) class.
- Recall
  Indicates the proportion of actual non-financial risk events that are correctly identified by the model. This metric reflects the system's ability to detect incidents early and minimize missed events.
- Recall@Top-10%
  Measures how many true events are captured within the top 10% of observations ranked by predicted risk. This is a business-aligned prioritization metric, reflecting real-world constraints where only a limited number of high-risk units can be reviewed or escalated.

Regression KPIs: (Severity Amount Prediction)

- MAE (Mean Absolute Error)
  Represents the average magnitude of prediction errors and provides an intuitive measure of typical severity estimation accuracy.
- RMSE (Root Mean Squared Error)
  Penalizes larger prediction errors more heavily, making it particularly relevant for operational risk severity modeling, where extreme losses can have disproportionate impact.

Explainability:

- SHAP Coverage and Plausibility
  Assesses whether model predictions can be consistently explained using SHAP values and whether the identified drivers align with domain knowledge (e.g., control effectiveness, operational stress, audit findings). This ensures transparency, auditability, and regulatory acceptability of the AI system.

Business KPIs

Beyond technical accuracy, the success of the system is evaluated using business-oriented KPIs that reflect its practical value to non-financial risk management.

Number of High-Impact Incidents Predicted in Advance

- Measures the system's effectiveness in identifying material risk events before they occur, enabling proactive intervention and mitigation.

Reduction in Unexpected Operational Losses

- Evaluates the downstream impact of improved risk prioritization and control actions, reflecting the system's contribution to reducing surprise losses and enhancing overall operational resilience.

This project aims to answer the questions:

- Which business units should be prioritized for improvement initiatives in Non-Financial Risk Management, based on the likelihood of experiencing a non-financial risk event within the next 30 days?
- What is the estimated amount of potential financial loss that could be avoided through proactive measures?

2. Data Collection & Understanding

Data Choice Set

A synthetic dataset was used to represent operational risk data while preserving confidentiality. The dataset contains daily snapshots for multiple business units, including KRIs, control metrics, audit indicators, HR stability metrics, IT change signals, and incident narratives.

This synthetic dataset simulates:

- real-world NFR data structure
- mixed data types
- event labels
- narratives for Natural Language Processing (NLP)
- enough volume for modelling

Due to the sensitive nature of operational loss data in banking, synthetic data was generated to realistically simulate patterns observed in real institutions while avoiding privacy and regulatory constraints.

Dataset Overview and Summary

The dataset used in this capstone project consists of 10,000 observations spanning a period of two years and covering 120 distinct business units within a banking context. Each record represents a time-based snapshot of operational conditions and control environments for a given unit, enabling both cross-sectional and temporal analysis of non-financial risk.

The dataset integrates a diverse set of variables capturing multiple dimensions of non-financial risk, including:

- Key Risk Indicators (KRIs), such as failed transaction counts, system downtime, queue length metrics, and historical incident frequency. These serve as early warning signals of operational stress.
- Risk and Control Self-Assessment (RCSA) measures, including the number of controls in place, average control effectiveness scores, and open audit findings. These provide insight into the strength of the control environment.
- Human resource (HR) metrics, such as employee attrition rates and average tenure. These reflect workforce stability and its potential impact on operational resilience.
- Information technology (IT) change indicators, including the occurrence of major system releases. These capture operational disruption and change-related risk.
- Incident narratives, consisting of free-text descriptions of past incidents offer qualitative context and potential enrichment for advanced analysis.
- External risk signals, such as regulatory alert scores and third-party or vendor incident counts. These represent exogenous risk factors beyond internal operations.
- Target variables, including a binary indicator of non-financial risk event occurrence within the next 30 days and a continuous variable representing the associated financial severity.

The dataset combines operational, control, human, technological, and external risk indicators to provide a holistic foundation for non-financial risk prediction and severity estimation.

Data Dictionary

The data dictionary below defines all variables used in this capstone project, including their data types and analytical roles. It provides a consistent reference for interpreting operational, control, and external risk indicators, as well as the target variables used for supervised learning, supporting transparency, reproducibility, and audit readiness.

Variable	Data Type	Description
unit_id	string	Business unit identifier (organizational entity under operational risk assessment)

Variable	Data Type	Description
date	string	Daily snapshot date of operational indicators ****convert to date during preprocessing*
kri_failed_txn_7d	integer	Count of failed customer or operational transactions logged over the last 7 days
kri_downtime_hours_30d	float	Total system downtime affecting the business unit over the past 30 days (in hours)
kri_queue_len_mean_14d	integer	Average operational backlog / queue length across workflows in the past 14 days
incidents_count_90d	integer	Number of recorded operational risk incidents or near-miss events in the past 90 days
sum_loss_amt_365d	float	Total monetary loss from operational incidents incurred in the last 12 months
days_since_last_incident	integer	Days elapsed since the most recent operational loss event or incident
num_controls	integer	Number of operational controls implemented in the business unit
avg_control_effectiveness	float	Average control performance rating (scale 1–5, where 5 is highly effective)
pct_controls_tested_12m	float	Percentage of controls tested or validated over the last 12 months
open_audit_finding_count	integer	Number of outstanding audit issues against the business unit
attrition_rate_90d	float	Percentage of staff leaving the business unit over the past 90 days
avg_tenure_months	float	Average employee tenure (in months) within the business unit
major_release_last_30d	integer	Whether a major IT change was released in the last 30 days (1 = yes; 0 = no)
open_changes_count	integer	Number of currently open change requests, release items, or configuration updates
incident_narrative	string	Narrative text describing past incidents, risk issues, or control breakdowns ****Change null values to "No narrative"*
regulatory_alert_score	float	Weighted score of regulatory alerts impacting the business unit (higher score = higher regulatory pressure)
third_party_incident_count	integer	Number of incidents attributed to vendors / suppliers in the past period
target_event_30d	integer	1 if a material operational loss event occurs within the next 30 days; else 0
target_severity_amt	float	Loss amount associated with the event occurring in the next 30 days (0 if no event)

Initial Data Observations

Upon reviewing the dataset, the following initial observations were made.

- The 'date' column is currently stored as an object type and will be converted to a datetime format for analysis.
- The only column with missing values is 'incident_narrative', with 8,792 out of 10,000 entries lacking data. This indicates that most records do not include an incident narrative.
    - o To address this, a new binary column will be created to indicate the presence of a narrative (1 = present, 0 = absent).
    - o Missing values in the 'incident_narrative' column will be replaced with 'No narrative'.
- The mean value of the 'target_event_30d' column is approximately 0.04. This means that about 4% of the records represent an operational risk incident occurring within the next 30 days. This highlights a highly imbalanced classification problem, which is typical in real-world operational risk and rare-event prediction scenarios.
- Most numerical features appear to have reasonable distributions. However, variables such as 'sum_loss_amt_365d' and 'target_severity_amt' display a wide range and are potentially skewed, with many zero values in 'target_severity_amt'. This is also expected in the context of NFR. Severity Amount is usually zero and when losses do occur, they tend to be highly skewed.

-------------

NOTE:

- Filename of the dataset is *synthetic_NFR_data.csv*
- Descriptive statistics is in *AI_Driven_NFR_Prediction_for_Banking.ipynb*. The whole notebook is designed and coded to be executable in Google Colab.

--------------

3. Data Preprocessing, Applied EDA & Feature Engineering

Data Cleaning and Preprocessing

Initial data cleaning steps were performed to ensure data consistency and analytical readiness.

- The date field was converted to a standard datetime format to support time-based analysis.
- The dataset was checked for duplicate records, and none were identified.
- To handle missing textual information, a binary indicator (has_incident_narrative) was created to capture the presence of an incident description. The missing values in the incident_narrative field were imputed with a neutral placeholder (“No narrative”). These steps ensured completeness while preserving potentially informative signals related to incident documentation.

- Potential outliers were identified but retained. These outliers represent critical rare events or extreme operational conditions relevant to risk modeling. Here are the reasons:
    - Failed Transactions: Detecting spikes is important. Removing them destroys early-warning signal.
    - Downtimes: Long downtimes are rare but high-impact. Central to operational risk modeling.
    - Queue Length: Extreme backlogs are real operational failures. Low outlier count. Thus, not distortionary.
    - Incident Counts: Removing them weakens signal.
    - Sum of Loss Amount: Operational losses are heavy-tailed by nature. Outliers are the risk.
    - % Controls Tested: Within bounds of 0-100%.
    - Open Audit Findings: Large counts indicate severe governance failure. High signal-to-noise for risk modeling.
    - Attrition rate: Within bounds of 0-100%.
    - Average Tenure: Extremely high tenure may reflect legacy teams while extremely low may mean starting to learn the operations.
    - Major Change Release: Binary and should not be treated as outliers.
    - Open Changes: High change volume correlates with failures. This is a valid risk driver.
    - Regulatory Alerts: Elevated regulatory signals are meaningful.
    - Third Party Incidents: Third-party incidents cluster heavily. Major driver of modern operational risk.

Applied EDA and Feature Engineering

Correlation Analysis, Relationship from Pair Plots, PCA

Strongest Correlations with Event Occurrence

- The high correlation with Severity Amount is expected, as an event occurring often leads to some severity. Open Audit Findings and Major Change Release show a positive correlation. It suggests that more audit findings and recent major releases are indicators of increased risk of an event.

Strongest Correlations with Severity Amount

- Similar to Event Occurrence, the occurrence of an event is a strong predictor of its severity. Higher Downtime and Queue Length (indicators of operational issues) also correlate positively with higher severity.

Inverse Correlations

- Average Control Effectiveness shows a slight inverse correlation with both target variables. This is logical in the NFR setting since higher control effectiveness should lead to fewer and less severe incidents.

Relationships from Pair Plots

- The pair plots for the top correlated features reveal various distribution patterns. For many features, the distribution appears skewed, especially for metrics like KRI Failure, Incident Counts, and Open Audit Findings. This indicates that lower values are more common, with occasional spikes. This aligns with the nature of risk events.

Separation by Event Occurrence

- While clear, distinct clusters are not immediately obvious in all plots, there are subtle differences in the distributions of features when colored by Event Occurrence. For instance, incidents (where target_event_30d is 1) tend to occur at higher values of KRI Failure, Downtime, Incident Count, Open Audit Finding, and Major Change Release, compared to non-incidents.

Severity Amount Distribution:

- The histograms for Severity Amount show a heavily skewed distribution, with most values at zero, and a long tail for non-zero values. This reflects that most days have no loss, but when losses occur, they can be substantial.

Clustering Tendencies from PCA

- The PCA plot, reducing the numerical features to two principal components, shows some degree of separation, particularly for target_event_30d = 1 (incident occurred). While not perfectly separable, data points corresponding to target_event_30d = 1 appear to form a somewhat distinct, albeit overlapping, region, particularly for higher values along Principal Component 1 and 2, which might be correlated with higher risk indicators.
- The explained variance ratios (PC1: 0.13, PC2: 0.09, Total: 0.22) indicate that the first two principal components capture only a small portion (22%) of the total variance in the dataset. This suggests that the dataset has high dimensionality and complexity, and the underlying structure is not easily captured by just two principal components. However, even with limited explained variance, some trends related to the target variable are visible.

Decisions for Feature Engineering for Non-Financial Risk

Feature engineering is used to turn raw data into meaningful risk signals that models can learn from and decision-makers can trust.

- Expose hidden risk patterns: As discussed in the previous section, raw indicators rarely show how risks interact. Feature engineering captures escalation effects that better reflect real-world risk.
- Improve predictive performance: Well-engineered features increase signal-to-noise, helping models detect rare events and estimate severity more accurately.
- Encode domain knowledge: Risk expertise is embedded directly into the data. This reduces reliance on purely statistical relationships.
- Support explainability and governance**:** Meaningful features make model outputs easier to interpret, justify, and defend to auditors, regulators, and risk committees.
- Handle data limitations**:** In rare-event and imbalanced settings, such as Event Occurrence, feature engineering compensates for limited positive examples by strengthening informative signals.

Correlation between Target Variables (Event Occurrence and Severity Amount)

- The values show high correlation between the two Target Variables. This means that Severity Amount is meaningful only when an Event Occurrence occurs. This is logically the nature of operational events and the dataset. No further feature engineering will be applied for this.

Correlations of other variables with Target Variables

- These show weak correlation overall. Hence, further feature engineering will be applied:
    - How much unresolved audit pressure exists relative to how strong the controls are?: audit_pressure = open_audit_finding_count / (avg_control_effectiveness + ε)
    - How stressed the change environment is during a major system release?: change_stress = major_release_last_30d × open_changes_count
    - How overloaded the operational process is?: ops_stress = kri_failed_txn_7d × kri_queue_len_mean_14d
    - How often incidents are occurring in an environment with weak controls?: control_gap = (1 − avg_control_effectiveness/5) × incidents_count_90d

Feature Distributions:

- In NFR, extremes or tails matter. Further feature engineering will be applied:
    - Is this business unit experiencing an unusually high volume of transaction failures?: high_failed_txn_flag = 1 if kri_failed_txn_7d > 90th percentile
    - Has the system experienced unusually severe or prolonged downtime recently?: high_downtime_flag = 1 if kri_downtime_hours_30d > 90th percentile
    - Is this business unit experiencing high incident counts? high_incidents_flag = 1 if incidents_count_90d > 90th percentile

PCA Clustering Tendencies

- Although PCA revealed subtle clustering of incident observations, the low variance explained indicates that operational risk is inherently high-dimensional. As such, dimensionality reduction will not be applied to the final models.

Feature Engineering

Feature engineering was applied as discussed in the previous section. The results are summarized below:

Correlation with Event Occurrence

- audit_pressure (0.041769) shows a slightly higher positive correlation than the original open_audit_finding_count (0.039673).
- change_stress (0.032233) has a notable positive correlation.
- control_gap (0.029536) also shows a positive correlation, slightly higher than individual incidents_count_90d.
- ops_stress (0.028020) demonstrates a positive correlation.
- high_failed_txn_flag (0.026546) shows a positive correlation.

Correlation with Severity Amount

- control_gap (0.051870) now shows the highest correlation (after target_event_30d) with severity, outperforming individual incidents_count_90d (0.031676) and avg_control_effectiveness (-0.037997) and highlighting its importance.
- ops_stress (0.043186) also exhibits a stronger correlation with severity compared to its constituent parts (kri_failed_txn_7d 0.028808, kri_queue_len_mean_14d 0.036233).
- audit_pressure (0.035837) also shows an improved correlation compared to open_audit_finding_count (0.029977).
- high_downtime_flag (0.024052) and high_failed_txn_flag (0.024210) also show positive correlations.

In summary, several of the engineered features, particularly control_gap, ops_stress, and audit_pressure, show improved or more meaningful correlations with the target variables compared to their individual components. This indicates that the feature engineering efforts have been successful in creating stronger signals for risk prediction.

Feature Importance & Explainability

Decision for Using SHAP

Global SHAP (SHapley Additive exPlanations) values was used to identify the primary drivers of operational risk. In a regulated banking context, an explainability tool must satisfy four non-negotiable requirements:

- Faithfulness – explanations must accurately reflect model behavior
- Consistency – important features must always be identified as important
- Local and global interpretability – explain individual predictions and overall model logic
- Audit and governance readiness – explanations must be defensible to auditors and regulators

SHAP is the only widely adopted explainability method that satisfies all four simultaneously.

While LIME, PDP, and ICE are also powerful explainability tools, they serve different purposes:

- LIME (Local Interpretable Model-agnostic Explanations): LIME excels at providing local explanations for individual predictions. It explains why a specific business unit was flagged as high-risk. While invaluable for detailed, instance-level scrutiny, it doesn't inherently provide the comprehensive global overview needed for feature selection across the entire dataset as effectively as SHAP.
- PDP (Partial Dependence Plots): PDPs show the average marginal effect of one or two features on the predicted outcome. They are great for understanding global trends and validating domain knowledge (e.g., how increasing downtime generally affects risk). However, they can mask heterogeneous relationships, as they average out individual effects.

- ICE (Individual Conditional Expectation) Plots: ICE plots disaggregate the average effect shown by PDPs, revealing how a feature affects each individual prediction. This is excellent for identifying specific edge cases or understanding a single unit's unique sensitivity to a feature. Like LIME, its strength lies in individual-level detail rather than comprehensive global ranking.

Summary of Results

The primary drivers of Non-Financial risk based on the SHAP analysis for both the Event Occurrence (classification) and Severity Amount (regression) prediction models are:

For Event Occurrence (Classification):

- target_severity_amt: Most influential, indicating that high potential severity significantly increases the likelihood of an event occurring.
- high_incidents_flag, incidents_count_90d: High historical incident counts are strong positive drivers.
- open_audit_finding_count, audit_pressure: More open audit findings or higher audit pressure increase event likelihood.
- high_failed_txn_flag, kri_failed_txn_7d: Elevated transaction failure rates are key indicators.
- avg_control_effectiveness: Lower control effectiveness suggests increased risk.
- regulatory_alert_score: Higher scores indicate higher event likelihood.
- sum_loss_amt_365d: Higher past loss amounts contribute to higher predicted likelihood.

For Severity Amount (Regression)

- target_event_30d: Overwhelmingly the most significant driver; an event predicted to occur (1) leads to a non-zero severity prediction.
- control_gap: A powerful predictor, directly correlating larger gaps with increased financial losses.
- ops_stress: A strong positive driver, indicating overloaded operational processes lead to costlier incidents.
- kri_downtime_hours_30d, high_downtime_flag: Higher downtime significantly drives up predicted severity.
- sum_loss_amt_365d: Higher historical loss amounts indicate a higher likelihood of future high severity events.
- avg_control_effectiveness: Lower control effectiveness contributes to higher predicted severity.
- audit_pressure: Higher audit pressure correlates with increased severity.
- regulatory_alert_score: A higher score increases predicted severity.

Data Analysis Key Findings

- Data for both classification and regression models were successfully prepared, including feature selection and an 80/20 train-test split (with stratification for classification).
  Both X_classification and X_regression datasets contain 24 features for 10,000 samples.

- A LightGBM Classifier was trained for Event Occurrence, addressing class imbalance by using scale_pos_weight set to 24.16.
- A LightGBM Regressor was trained for severity prediction, utilizing the regression_l1 objective for robustness to outliers.
- SHAP analysis confirmed the strong interdependence of the two target variables: target_severity_amt was the most influential predictor for Event Occurrence, and target_event_30d was the most significant driver for severity prediction.
- Several engineered features emerged as highly important for both models, including control_gap, ops_stress, audit_pressure, high_incidents_flag, high_downtime_flag, and high_failed_txn_flag, validating the feature engineering efforts.
- Consistent risk indicators were identified across both models, notably those related to operational stress (e.g., failed transactions, downtime), control weaknesses (e.g., audit findings, control effectiveness), historical incident activity, and regulatory signals.

Observations and Actions

The following results are highlighted:

- For Event Occurrence (Classification): target_severity_amt is the most influential, indicating that high potential severity significantly increases the likelihood of an event occurring.
- For Severity Amount (Regression): target_event_30d is overwhelmingly the most significant driver; an event predicted to occur (1) leads to a non-zero severity prediction.

When target_event_30d = 1, then target_severity_amt > 0.

- This means that target_event_30d and target_severity_amt are logically dependent.
- Including one as a feature to predict the other gives the model future information
- This violates a fundamental rule: A target variable (or a proxy of it) must never appear as a feature.
- This signifies target leakage.

Action to Address Leakage

- Apply SHAP to identify the primary drivers of Non-Financial risk excluding target_severity_amt when training classification model and excluding target_event_30d when training regression model. This is to address target leakage.

Summary of Results After Addressing Leakage

The primary drivers of Non-Financial risk based on the SHAP analysis for both the Event Occurrence (classification) and Severity Amount (regression) prediction models are:

For Event Occurrence (Classification):

- high_incidents_flag, incidents_count_90d: High historical incident counts are strong positive drivers.

- open_audit_finding_count, audit_pressure: More open audit findings or higher audit pressure increase event likelihood.
- high_failed_txn_flag, kri_failed_txn_7d: Elevated transaction failure rates are key indicators.
- avg_control_effectiveness: Lower control effectiveness suggests increased risk.
- regulatory_alert_score: Higher scores indicate higher event likelihood.
- sum_loss_amt_365d: Higher past loss amounts contribute to higher predicted likelihood.

For Severity Amount (Regression):

- control_gap: A powerful predictor, directly correlating larger gaps with increased financial losses.
- ops_stress: A strong positive driver, indicating overloaded operational processes lead to costlier incidents.
- kri_downtime_hours_30d, high_downtime_flag: Higher downtime significantly drives up predicted severity.
- sum_loss_amt_365d: Higher historical loss amounts indicate a higher likelihood of future high severity events.
- avg_control_effectiveness: Lower control effectiveness contributes to higher predicted severity.
- audit_pressure: Higher audit pressure correlates with increased severity.
- regulatory_alert_score: A higher score increases predicted severity.

Data Analysis Key Findings

- Data for both classification and regression models were successfully prepared, including feature selection and an 80/20 train-test split (with stratification for classification). X_classification and y_classification contain 24 features for 10,000 samples. The regression dataset was filtered to include only 397 instances where target_severity_amt > 0, and then split into training and testing sets.
- A LightGBM Classifier was re-trained for Event Occurrence, addressing class imbalance by using scale_pos_weight set to 24.16.
- A LightGBM Regressor was re-trained for severity prediction, utilizing the regression_l1 objective for robustness to outliers, on the filtered dataset.
- The target leakage was successfully addressed by excluding target_severity_amt from classification features and target_event_30d from regression features.
- SHAP analysis for the corrected models revealed that several engineered features (control_gap, ops_stress, audit_pressure, high_incidents_flag, high_downtime_flag, and high_failed_txn_flag) emerged as highly important for both models, validating the feature engineering efforts.
- Consistent risk indicators were identified across both models, notably those related to operational stress (e.g., failed transactions, downtime), control weaknesses (e.g., audit findings, control effectiveness), historical incident activity, and regulatory signals.

Feature Selection

- Feature selection was performed using an embedded approach. Tree-based gradient boosting inherently selects informative features during training by prioritizing splits with positive gain, allowing the model to capture non-linear relationships and feature interactions. SHAP values were subsequently used to validate and interpret feature importance, confirming that engineered and NFR-relevant features contributed meaningfully to model predictions.
- Filter methods were not relied upon as they assess features in isolation and may overlook interaction effects that are central to operational risk.
- Wrapper methods were also not selected due to their computational cost and limited added value in this context, particularly given the strong built-in feature selection capabilities and interpretability of tree-based models.

NOTE:

- All detailed steps for preprocessing, EDA, visual representations (distributions, correlation matrix, tables, pair plots, etc.) are documented with reproducible code and can be found in *AI_Driven_NFR_Prediction_for_Banking.ipynb*. The whole notebook is designed and coded to be executable in Google Colab.

----

4. Model Implementation

Supervised Models

Multiple supervised learning models were explored for both Classification and Regression.

- LightGBM (existing model)
- Logistic Regression
- Random Forest
- Decision Tree
- SVM (Classification only)

Compare the Classification Models using the metrics:

- ROC-AUC
- PR-AUC
- Recall
- Recall@Top-10%

Compare the Regression Models using the metrics:

- RMSE
- MAE

Comparing the Classification Models

The table below shows the performance of the different Classification Models (for Event Occurrence):

index	ROC-AUC	PR-AUC	Recall	Recall@Top-10%
LightGBM	0.49443526	0.04035478	0.01265823	0.13924051
Logistic Regression	0.52948425	0.04168236	0.36708861	0.08860759
Random Forest	0.47994518	0.03739842	0	0.07594937
Decision Tree	0.49582232	0.0392383	0.03797468	0.08860759
SVC	0.46668072	0.03611753	0.15189873	0.06329114

Overall Performance

- The classification task aims to predict whether a non-financial risk event (Event Occurrence) will occur within the next 30 days. As established during exploratory analysis, the dataset is highly imbalanced**,** with approximately 4% positive events. In such rare-event settings, conventional metrics such as ROC-AUC provide limited insight, while Recall, Precision–Recall AUC, and particularly Recall@Top-10% are more aligned with operational risk management objectives.
- Across all evaluated models, ROC-AUC values are close to 0.5 and PR-AUC values are low, which is expected given the rarity of events and the weak marginal signal of individual risk indicators. These results do not indicate model failure but instead reflect the inherent difficulty of predicting rare non-financial risk events.

Model-Specific Interpretation

- LightGBM: LightGBM achieved a ROC-AUC of approximately 0.49 and low overall recall at the default threshold, indicating a conservative classification tendency. However, it achieved the highest Recall@Top-10% (≈ 13.9%) among all models, meaning it captures the largest proportion of true risk events when attention is restricted to the highest-risk decile. This aligns well with real-world NFR practices, where risk teams prioritize a limited subset of units for review rather than acting on all alerts.
- Logistic Regression: Logistic Regression produced the highest ROC-AUC (≈ 0.53) and the strongest overall recall (≈ 36.7%), demonstrating its effectiveness as a baseline model. However, its Recall@Top-10% (≈ 8.9%**)** is lower than LightGBM's, suggesting that while it identifies more events overall, it is less effective at ranking risk within the most constrained review capacity. This limits its usefulness for prioritization-driven decision-making.
- Random Forest: Random Forest achieved near-random ROC-AUC and zero recall at the default threshold, indicating poor sensitivity to rare events. Its Recall@Top-10% (**≈** 7.6%) suggests limited ranking capability compared to LightGBM. This reflects the tendency of Random Forests to underperform in highly imbalanced datasets without extensive tuning.
- Decision Tree: The single Decision Tree model showed weak performance across all metrics, including low recall and modest Recall@Top-10%. While inherently

interpretable, its instability and poor generalization make it unsuitable as a final model for non-financial risk prediction.

- Support Vector Classifier (SVC): SVC demonstrated low ROC-AUC and PR-AUC values, with moderate recall but the lowest Recall@Top-10% among the models. This confirms that margin-based classifiers struggle with severe class imbalance and lack the ranking effectiveness required for operational risk prioritization.

Observations and Decisions

- The results indicate that risk ranking performance is more informative than raw classification accuracy in this capstone. While Logistic Regression performs well as a baseline in terms of overall recall**,** LightGBM provides superior prioritization capability, capturing the highest proportion of true risk events within the top-risk decile. This makes LightGBM more suitable for real-world non-financial risk management, where investigative resources are limited and explainable prioritization is essential.
- Although overall classification metrics are modest due to severe class imbalance, LightGBM demonstrates superior Recall@Top-10%, indicating stronger risk ranking capability. This aligns with non-financial risk management objectives, where identifying the most critical high-risk units is more valuable than maximizing overall recall.
- Improve Classification Models: Given the generally poor performance of all classification models (ROC-AUC around 0.5, low PR-AUC and Recall), it is crucial to investigate potential issues such as severe class imbalance (despite using class_weight='balanced'), feature engineering, or the inherent predictability of the target variable. Further hyperparameter tuning and exploring more advanced techniques like anomaly. In this case, use tuning threshold. The default 0.5 threshold is not correct.

Comparing the Classification Models After Tuning

The table below shows the performance of the different Classification Models after tuning (for Event Occurrence):

index	Optimal Recall Threshold	Optimal Recall	Optimal Recall@Top-10% Threshold	Optimal Recall@Top-10%
LightGBM	0.05	0.63291139	0.05	0.13924051
Logistic Regression	0.05	1	0.05	0.08860759
Random Forest	0.05	0.27848101	0.05	0.07594937
Decision Tree	0.05	0.03797468	0.05	0.08860759
SVC	0.05	0.01265823	0.05	0.06329114

Context and Rationale for Threshold Tuning

- The classification task involves predicting rare non-financial risk events, with approximately 4% positive cases in the dataset. In such highly imbalanced settings, the default probability threshold of 0.5 is inappropriate because leads to extremely low sensitivity. Threshold tuning was therefore applied to shift the decision boundary toward improved event detection and to align model outputs with risk prioritization objectives rather than strict classification accuracy.

- All models were evaluated at an optimized threshold of 0.05. This reflects a deliberate trade-off that favors recall over precision in line with operational risk practices.

Model-Specific Interpretation After Threshold Tuning

- LightGBM: LightGBM achieved a substantial improvement in recall, reaching approximately 63.3%, while also delivering the highest Recall@Top-10% (≈ 13.9%) among all models. This indicates that LightGBM is effective not only at identifying a large proportion of true risk events but also at concentrating those events within the highest-risk decile, which is critical when review capacity is constrained. This balanced improvement highlights LightGBM's strength in rare-event ranking and prioritization.
- Logistic Regression: Logistic Regression achieved perfect recall (100%) at the tuned threshold, indicating that all true events were flagged. However, its Recall@Top-10% (≈ 8.9%) is notably lower than LightGBM's. This suggests that the model achieves high recall by broadly flagging observations as high risk, resulting in weaker prioritization and reduced practical usefulness when resources are limited.
- Random Forest: Random Forest showed a moderate recall of approximately 27.8%, but its Recall@Top-10% remained relatively low (≈ 7.6%). This indicates limited improvement in risk concentration despite threshold tuning, reinforcing its weaker suitability for rare-event prioritization in this context.
- Decision Tree: The Decision Tree model continued to exhibit very low recall and modest Recall@Top-10%, confirming its instability and limited generalization capability even after threshold adjustment.
- Support Vector Classifier (SVC): SVC remained largely insensitive to threshold tuning, with near-zero recall and the lowest Recall@Top-10%. This further demonstrates the model's poor alignment with severely imbalanced, tabular NFR data.

Observations and Decisions

- Threshold tuning significantly improved the ability of models to detect rare non-financial risk events. However, improved recall alone is insufficient for operational risk management. The most relevant metric is Recall@Top-10%**,** which reflects the ability to prioritize risk within constrained investigative capacity. LightGBM consistently outperforms other models on this metric, indicating superior risk ranking and practical decision support.
- While Logistic Regression achieved perfect recall after threshold tuning, this came at the expense of meaningful risk prioritization. LightGBM provides the best balance between high event detection and effective concentration of true events within the top-risk segment, making it the most suitable model for non-financial risk management.

Comparing the Regression Models

The table below shows the performance of the different Regression Models (for Severity Amount):

index	RMSE	MAE
LightGBM	12907.0186	7765.67504
Linear Regression	12310.5532	8256.58213
Random Forest	12797.7757	8202.28299
Decision Tree	15250.3297	10554.4323

Overall Performance

- The regression task estimates the financial severity of non-financial risk events (Severity Amount), conditional on an event occurring. Loss severity data in operational risk is typically highly skewed and heavy-tailed, with a small number of extreme losses driving overall risk exposure. As a result, model evaluation focuses on both Mean Absolute Error (MAE), which reflects typical prediction error, and Root Mean Squared Error (RMSE), which penalizes large deviations more heavily.

Model-Specific Interpretation

- LightGBM: LightGBM achieved an RMSE of approximately 12,907 and the lowest MAE (≈ 7,766) among all evaluated models. The lower MAE indicates stronger performance in estimating the typical severity of loss events. Its ability to model non-linear relationships and feature interactions makes it well-suited to capturing escalation effects in operational risk, even though its RMSE is slightly higher than that of Linear Regression.
- Linear Regression: Linear Regression produced the lowest RMSE (≈ 12,311) but a higher MAE **(**≈ 8,257**)**. This suggests that while the model fits average loss levels reasonably well, it is less accurate for the majority of observations and may underperform when capturing non-linear drivers of severity. Given the complex and interaction-driven nature of non-financial risk losses, this limits its suitability as a final model despite competitive RMSE.
- Random Forest: Random Forest achieved intermediate performance, with RMSE and MAE values falling between Linear Regression and LightGBM. While it can model non-linear relationships, its performance does not exceed that of LightGBM, and it offers less stable and less transparent explainability for severity drivers.
- Decision Tree: The Decision Tree model performed worst across both metrics, exhibiting substantially higher RMSE and MAE. This reflects its tendency to overfit and its inability to generalize effectively in sparse, noisy loss severity data.

Observations and Decisions

- Although Linear Regression achieves the lowest RMSE, LightGBM provides the best balance between accuracy and robustness, as evidenced by its lowest MAE and superior ability to capture non-linear severity drivers. In the context of non-financial risk management—where understanding and explaining the drivers of loss magnitude is critical—LightGBM is therefore the most appropriate severity model.
- While Linear Regression marginally outperforms in RMSE, LightGBM achieves the lowest MAE and offers greater robustness and interpretability for non-linear loss dynamics, making it the preferred model for estimating non-financial risk severity.

Unsupervised Models

Although Supervised learning models are the appropriate models to use for the NFR problem, multiple unsupervised learning models will be explored for academic purposes:

- K-Means
- DBSCAN
- Hierarchical (Elbow, Silhouette)

Metrics:

- Silhouette Score
- PCA-based visual inspection

Comparing the Clustering Models

- Three unsupervised clustering techniques—K-Means, DBSCAN, and Hierarchical (Agglomerative) Clustering—were evaluated to explore latent structure within the non-financial risk (NFR) feature space. Model suitability was assessed using Silhouette Scores and PCA-based visual inspection.
- K-Means clustering identified an optimal solution of three clusters, supported by both the Elbow Method and Silhouette analysis, achieving a Silhouette Score of approximately 0.214. The PCA-reduced visualization showed moderately separated clusters with some overlap, indicating the presence of distinguishable but not sharply separated risk groupings.
- DBSCAN identified four clusters while classifying a substantial number of observations (1,981) as noise. Its Silhouette Score of approximately 0.183 (excluding noise points) was lower than the other approaches. While DBSCAN proved effective at highlighting dense core regions and isolating anomalous observations, the high proportion of noise suggests that much of the NFR data does not form tightly packed clusters, limiting its usefulness for broad risk segmentation.
- Hierarchical (Agglomerative) Clustering also indicated three clusters as optimal and achieved the highest Silhouette Score (≈ 0.239) among all tested methods. The PCA visualization showed reasonably well-defined clusters, comparable to K-Means but with slightly improved separation. This suggests a clearer underlying structure when hierarchical relationships between observations are considered.

Suitability for Non-Financial Risk Management

- Among the evaluated methods, Hierarchical Clustering (Agglomerative) appears most suitable for identifying latent patterns in the NFR dataset. Its higher Silhouette Score and balanced cluster structure indicate better internal cohesion and separation, which is desirable for exploratory segmentation of risk profiles. K-Means performs comparably and serves as a strong baseline, while DBSCAN is better positioned as a supplementary technique for outlier and anomaly identification rather than primary clustering.
- Importantly, the relatively modest Silhouette Scores across all methods highlight that NFR data does not naturally form strongly separable clusters, reflecting the complex, overlapping, and multi-factor nature of operational and non-financial risks. This reinforces the appropriateness of using unsupervised clustering as an exploratory and complementary analysis, rather than as a core predictive mechanism.

Recommender System

In the non-financial risk setting in banking, there are no user-preference data. There are no user–item interaction histories (required for collaborative filtering). Risk prioritization must be explainable and auditable. Content-based methods can justify recommendations using risk attributes (KRIs, controls, audit findings). Hence, content-based recommendation is appropriate because risk prioritization should be driven by unit risk characteristics and not peer behavior.

Data Preparation and Risk Profiling

- To support a content-based recommendation approach, a subset of 24 relevant numerical risk indicators was selected from the original dataset, excluding identifiers and target variables to prevent leakage. These features were aggregated at the business unit level (unit_id) using mean values, resulting in 120 distinct unit-level risk profiles, each represented by a consistent set of operational, control, and governance attributes. The aggregated profiles were subsequently scaled using StandardScaler to ensure comparability across features and to prevent dominance by variables with larger numeric ranges.

Similarity Computation

- Cosine similarity was selected as the similarity measure due to its suitability for comparing high-dimensional feature vectors and its interpretability in terms of relative risk profile alignment. Applying cosine similarity to the scaled unit profiles produced a 120 × 120 similarity matrix, capturing the degree of resemblance between all pairs of business units based on their non-financial risk characteristics.

Recommendation System Implementation and Validation

- A recommendation function was implemented to retrieve the top-N most similar business units for a given query unit, excluding the unit itself to avoid trivial matches. Testing the system using UNIT_001 successfully returned a ranked list of five peer units with closely aligned risk profiles. This demonstrates the system’s ability to identify peer units exhibiting similar operational and control risk patterns, supporting comparative risk analysis and targeted oversight.

Suitability for Non-Financial Risk Management

- The content-based recommendation approach is well-suited to the non-financial risk domain. It relies exclusively on observable risk attributes rather than historical user-interaction data, which is typically unavailable in NFR contexts. Furthermore, recommendations are fully explainable and auditable, as similarities can be traced directly to shared risk drivers such as control effectiveness, incident history, and operational stress indicators. This makes the approach appropriate for governance-driven environments and supports proactive risk management interventions.

Area for Further Study for Non-Financial Risk Management

- Incorporate Textual Data: Leverage the incident_narrative column using Natural Language Processing (NLP) techniques to extract thematic risk categories or sentiment.

This 'unstructured content' could enrich unit profiles and capture nuanced similarities not apparent in numerical metrics.

- Explore Different Aggregation Methods: Instead of just the mean, consider other aggregation functions (e.g., median, maximum, standard deviation, time-weighted averages) for different features or for different time windows. For instance, the maximum KRI value in a period might be more indicative of risk than the average.
- Dynamic Profiles: Implement a system where unit profiles are updated more frequently (e.g., weekly or monthly) to reflect evolving risk landscapes, rather than static aggregation over the entire historical period. This would enable more timely and relevant recommendations.
- Hybrid Similarity Metrics: Experiment with other similarity metrics, potentially weighting them differently based on domain expertise, or creating a hybrid metric that combines aspects of different measures.
- Contextual Recommendations: Introduce contextual factors (e.g., business line, geographic location) into the similarity calculation. Units might be similar in risk profile, but recommendations could be refined to prioritize similar units within the same business line for more actionable insights.
- Anomaly Detection for Recommendations: Instead of just recommending 'similar' units, the system could identify units that are dissimilar to a healthy baseline or to their peer group, flagging them as potential emerging risks.

Deep Learning

Deep learning models were not performed due to limited data volume, tabular feature structure, and the need for transparent, auditable explanations in a non-financial risk context.

Area for Further Study for Non-Financial Risk Management

- Deep learning techniques are most appropriate for non-financial risk applications involving unstructured or sequential data, such as incident narratives, regulatory communications, and time-series escalation patterns.

NOTE:

- All detailed steps for pre-processing, model creation and training, visuals, tables, and graphs are documented with reproducible code and can be found in *AI_Driven_NFR_Prediction_for_Banking.ipynb*. The whole notebook is designed and coded to be executable in Google Colab.
- Classification Models: lightgbm_classifier.joblib, logistic_regression_classifier.joblib, random_forest_classifier.joblib, decision_tree_classifier.joblib, svc_classifier.joblib
- Regression (Severity) Models: lightgbm_regressor.joblib, linear_regression_regressor.joblib, random_forest_regressor.joblib, decision_tree_regressor.joblib
- Unsupervised / Auxiliary Models: kmeans_unsupervised.joblib, hierarchical_unsupervised.joblib, dbscan_unsupervised.joblib
- Recommender: cosine_sim_matrix.joblib

- Preprocessing / Feature Artifacts: scaler_cf.joblib (Feature scaler used in preprocessing), unit_profiles_df.joblib (Precomputed unit-level profiles (used for aggregation / recommendation logic))

----

5. Critical Thinking → Ethical AI & Bias Auditing

Summary of Decisions

This section summarizes the decisions made for this project. For the detailed rationale and thought process behind the decisions, kindly refer to the corresponding previous sections.

Data Pre-Processing and Feature Engineering

- The date field was converted to a standard datetime format to support time-based analysis.
- Potential outliers were identified but retained. These outliers represent critical rare events or extreme operational conditions relevant to risk modeling.
- Correlations of other variables with Target Variables show weak correlation overall. Further feature engineering was applied.
- Feature Distributions: In NFR, extremes or tails matter. Further feature engineering Further feature engineering was applied.
- PCA Clustering Tendencies: The low variance indicates that operational risk is inherently high-dimensional. As such, dimensionality reduction was not applied to the final models.

Explainability and Feature Selection

- SHAP values was used to identify the primary drivers of operational risk. In a regulated banking context, an explainability tool must satisfy non-negotiable requirements: Faithfulness, Consistency, Local and global interpretability, Audit and governance readiness.
- Embedded approach for feature selection was used because this allows the model to capture non-linear relationships and feature interactions.

Model Selection

- Supervised Learning Models: Best choice among the different types of learning models because it directly models known risk outcomes (i.e. Event Occurrence and Severity Amount)
    - Classification: LightGBM (tuned) provides the best balance between high event detection and effective concentration of true events within the top-risk segment.
    - Regression: LightGBM achieves the lowest MAE and offers greater robustness and interpretability for non-linear loss dynamics.

- Unsupervised Learning Models: Unsupervised clustering must be used as an exploratory and complementary analysis, rather than as a core predictive mechanism. Based on metrics, Hierarchical Clustering (Agglomerative) appears most suitable for identifying latent patterns in the NFR dataset.
- Recommender: Content-based recommendation approach is well-suited to the non-financial risk domain rather than Collaborative (user-based).
- Deep Learning: Not performed due to limited data volume, tabular feature structure, and the need for transparent, auditable explanations in a non-financial risk context. This may be used for unstructured or sequential data, such as incident narratives, regulatory communications, and time-series escalation patterns.

Addressing Limitations

Class Imbalance (for Classification Model)

- The Problem:
  - The Target Variable Event Occurrence was highly imbalanced, with only about 4% of records representing an actual event. If not addressed, models tend to classify most instances as the majority class (no event), leading to poor performance in detecting the rare positive class.
- How Was It Addressed:
  - Stratified Sampling: When splitting the data into training and testing sets for classification, stratify=y_classification was used. This ensured that both the training and test sets maintained the same proportion of positive and negative classes as the original dataset.
  - Algorithm-level Weighting (scale_pos_weight): For the LightGBM Classifier, scale_pos_weight was calculated and applied. This parameter tells the model to give more importance to the minority class during training, effectively penalizing misclassifications of the rare event more heavily. The calculated scale_pos_weight was approximately 24.16.
  - Threshold Tuning: Recognizing that a default classification threshold of 0.5 is inappropriate for highly imbalanced datasets, extensive threshold tuning was performed. Various thresholds (from 0.05 to 0.49) were evaluated to optimize for business-critical metrics like Recall and Recall@Top-10%. This allowed effectively flagging more true positive events, even if it meant accepting a slightly higher number of false positives.

Target Leakage

- The Problem:
  - Target leakage occurs when information about the target variable, which would not be available at the time of prediction, is included in the features. This can lead to overly optimistic model performance that doesn't generalize to new, unseen data. This was the case for the highly positive correlation between the two Target Variables (Severity Amount and Event Occurrence).
- How Was It Addressed:
  - Exclusion from Classification Features: Initially, target_severity_amt (Severity Amount) showed a very high correlation with target_event_30d (Event

Occurrence). Since the severity of an event would only be known after the event occurs, including it to predict event occurrence would be leakage. Target_severity_amt was explicitly excluded from the feature set used to train the classification model.

    - Exclusion from Regression Features: Similarly, target_event_30d (Event Occurrence would be known before estimating its severity. However, for predicting target_severity_amt (which is often 0 when no event occurs), including target_event_30d would be problematic. Target_event_30d was explicitly from the feature set used to train the regression model. Regression dataset was also filtered to only include instances where target_severity_amt > 0. This ensures learning to predict severity when an event actually happened.

Overfitting

- The Problem:
    - Overfitting occurs when a model learns the training data too well, capturing noise and specific patterns that don't generalize to new data. This results in excellent performance on training data but poor performance on unseen test data.
- How Was It Addressed:
    - Train-Test Split: An 80/20 train-test split was consistently used with random_state=42 across all supervised models. This provided an unseen dataset for objective model evaluation. This ensured that the reported metrics reflect the model's generalization capability rather than just its ability to memorize training data.
    - Tree-Based Models (LightGBM, Random Forest, Decision Tree): While decision trees can easily overfit, ensemble methods like LightGBM and Random Forest inherently mitigate this. LightGBM, a gradient boosting machine, uses regularization techniques to control tree complexity and prevent overfitting. Random Forests, by building multiple trees on bootstrapped samples and averaging their predictions, also reduce variance and overfitting.
    - Regression Model Behavior: For the LightGBM Regressor, warnings about "no further splits with positive gain" were observed. This indicates that the model was conservatively stopping tree growth when it determined that further splits would not significantly improve performance on the training data, effectively preventing it from overfitting to sparse or noisy patterns in the severity data.
    - MAE Objective for Regression: By using objective='regression_l1' (Mean Absolute Error) for the regression model, a metric that is more robust to outliers than Mean Squared Error (MSE) is better. This helps prevent the model from excessively adjusting to extreme values, which could be noise or rare anomalies, thereby reducing overfitting to such points. The aim is to build robust and reliable models that could provide actionable insights for non-financial risk management by systematically applying these techniques.

Bias Detection & Fairness Auditing

Absence of Direct Sensitive Features in Synthetic Data:

- Based on the review of the df.columns.tolist() output, the provided synthetic dataset does not contain any direct sensitive features such as gender, age, ethnicity, or socioeconomic status. Therefore, direct quantitative bias detection or fairness auditing on these explicit demographic attributes is not feasible within this synthetic context.

Conceptual Discussion of Potential Biases in a Real-World NFR Scenario:
If this were real-world banking data, potential sources of bias could arise, even without explicit demographic features:

- unit_id as a Proxy: The unit_id feature, while an anonymized identifier here, could in a real scenario represent business units tied to specific geographical regions, product lines, or customer segments that might disproportionately serve certain demographic groups. For example, a unit heavily serving a low-income area might consistently show higher kri_failed_txn_7d or sum_loss_amt_365d not due to inherent unit inefficiency, but due to systemic economic disparities affecting its customer base.
- HR Metrics (attrition_rate_90d, avg_tenure_months): These metrics could implicitly correlate with age, experience levels, or even indirectly with gender or socioeconomic background if certain roles or locations tend to have specific demographic profiles. Bias could manifest if the model penalizes units with higher attrition rates that are, for instance, predominantly staffed by younger employees or specific demographic groups.
- incident_narrative: If free-text narratives contained biased language or focused more on incidents in certain types of units due to historical reporting practices, this could introduce bias that NLP models might pick up and amplify.
- Historical Data Bias: Even if features aren't direct proxies, historical operational data can reflect past human biases in decision-making, resource allocation, or even fraud detection, leading to models that perpetuate these biases.

Key Fairness Metrics for a Real-World Banking Context:
In a real-world setting and for a comprehensive fairness audit of the LightGBM models, evaluation of metrics across identified sensitive groups (or their proxies) should be conducted:

- For LightGBM Classification Model (Event Occurrence prediction)**:**
  - Statistical Parity Difference (SPD): Measures the difference in the positive prediction rate between the protected and unprotected groups (e.g., P(Y_pred=1 | A=protected) - P(Y_pred=1 | A=unprotected)).
  - Equal Opportunity Difference (EOD): Focuses on the difference in true positive rates (Recall) between groups (e.g., P(Y_pred=1 | Y_true=1, A=protected) - P(Y_pred=1 | Y_true=1, A=unprotected)). Critical for ensuring that the model correctly identifies actual risk events equally well across different groups.
  - Average Odds Difference (AOD): Averages the absolute differences in false positive rates and true positive rates across groups.
  - Predictive Equality Difference: Difference in false positive rates (P(Y_pred=1 | Y_true=0, A=protected) - P(Y_pred=1 | Y_true=0, A=unprotected)).
- For LightGBM Regression Model (Severity Amount prediction)
  - Mean Absolute Error (MAE) Difference: Compares the MAE of predictions for the protected group versus the unprotected group. (e.g., MAE(Y_true, Y_pred | A=protected) - MAE(Y_true, Y_pred | A=unprotected)).

    - Root Mean Squared Error (RMSE) Difference: Similar to MAE difference, but for RMSE.
    - Over/Under-prediction Bias: Examines if the model systematically over-predicts or under-predicts severity for certain groups. For example, calculating E[Y_pred - Y_true | A=protected] vs. E[Y_pred - Y_true | A=unprotected].

Proposed Mitigation Strategies:
If conceptual biases were identified or if the system is deployed with real data, the following mitigation strategies are recommended:

- Data Debiasing: Techniques like re-sampling (e.g., reweighing observations in the training data to balance group representation), or counterfactual data augmentation to reduce undesirable correlations.
- Model-Agnostic Fairness Algorithms: Applying pre-processing (e.g., Reweighing, Disparate Impact Remover), in-processing (e.g., Adversarial Debiasing), or post-processing (e.g., Reject Option Classification, Calibrated Equalized Odds) methods from fairness toolkits (e.g., AIF360, Fairlearn).
- Threshold Adjustment: Adjusting classification thresholds for different sensitive groups to achieve equal opportunity or other relevant fairness criteria, especially critical given the finding that threshold tuning significantly impacts recall.
- Feature Engineering with Fairness in Mind: Carefully creating or modifying features to ensure they do not inadvertently encode or amplify biases. Removing proxies for sensitive attributes if they are found to contribute to unfair outcomes.
- Continuous Monitoring: Implementing feedback loops and regular audits to detect emerging biases as data distributions change over time.

Ethical Implications for the Banking Context:
Addressing bias and ensuring fairness in AI systems for banking is paramount due to several ethical implications:

- Regulatory Compliance: Banking is a heavily regulated industry. Biased models can lead to non-compliance with anti-discrimination laws (e.g., fair lending laws, consumer protection regulations), resulting in severe penalties, fines, and legal action.
- Reputational Risk: Perceived or actual bias in risk prediction can severely damage a bank's reputation, leading to loss of customer trust, negative public perception, and reduced market share.
- Impact on Customers/Business Units: If a model unfairly flags certain business units as high-risk, it could lead to misallocation of resources, unfair performance evaluations, punitive measures, or even job losses for individuals within those units. Conversely, if it under-predicts risk for other groups, it could lead to unforeseen losses.
- Financial Inclusion and Equity: Biased risk models can perpetuate existing inequalities, potentially denying fair access to credit, services, or opportunities for specific customer segments, hindering financial inclusion efforts.
- Responsible AI Development: Banks have an ethical responsibility to develop and deploy AI systems that are fair, transparent, and accountable, upholding principles of justice and avoiding harm to individuals or groups.

In conclusion, while the synthetic nature of the dataset precludes direct bias detection, a real-world implementation would necessitate rigorous fairness auditing, proactive bias mitigation strategies, and a strong ethical framework to ensure the AI-driven NFR system operates responsibly and equitably.

Solving the Business Problem

This project aims to answer the following questions:

- Which business units should be prioritized for improvement initiatives in Non-Financial Risk Management, based on the likelihood of experiencing a non-financial risk event within the next 30 days?
- What is the estimated amount of potential financial loss that could be avoided through proactive measures?

Priority Business Units:

- These are the business units that will most-likely experience a non-financial risk event in the next 30 days (Event Occurrence).
- Using the LightGBM model (classifier with tuned threshold) to the predicted probabilities of the units that will most-likely experience an event in the next 30 days. The results show:

index	unit_id	predicted_proba
115	UNIT_116	0.97253594
111	UNIT_112	0.96451074
44	UNIT_045	0.9642574
97	UNIT_098	0.96333684
69	UNIT_070	0.96265765
50	UNIT_051	0.96080203
89	UNIT_090	0.95926361
16	UNIT_017	0.9590591
31	UNIT_032	0.95791982
102	UNIT_103	0.95666425

- These are the business units that should be prioritized for improvement initiatives in Non-Financial Risk Management.

Avoidable Potential Financial Loss:

- This is the total amount of mitigated potential financial loss coming from the top 10 priority units.
- Using the LightGBM model (regressor), the total predicted potential financial loss that could be avoided through proactive measures is $7,620,235.78.

Future Recommendations and Areas for Further Study

Integrated Risk Assessment Across Frequency and Severity

- The observed strong relationship between event occurrence and loss severity highlights the importance of treating non-financial risk as a unified problem rather than isolated predictive tasks. Future implementations should further integrate frequency and severity modeling outputs into a consolidated risk scoring framework to support holistic decision-making and prioritization.

Continued Emphasis on Domain-Driven Feature Engineering

- The significant performance and interpretability gains achieved through domain-specific engineered features demonstrate the critical role of subject-matter expertise in non-financial risk modeling. Future research could expand this approach by incorporating additional interaction features, stress indicators, and escalation metrics derived from operational workflows and control processes.

Dynamic and Adaptive Thresholding Strategies

- Threshold tuning proved essential for transforming model outputs into actionable insights under severe class imbalance. Future work could explore dynamic or adaptive thresholding mechanisms that adjust based on changing risk appetite, operational capacity, or external risk conditions, enabling more responsive and context-aware risk prioritization.

Refinement of Business-Aligned Evaluation Metrics

- The effectiveness of Recall and Recall@Top-K% underscores the need to move beyond traditional classification metrics in NFR contexts. Further studies could formalize these metrics within risk governance frameworks and explore cost-sensitive evaluation approaches that explicitly model the trade-offs between missed events and investigation effort.

Incorporation of Unstructured Data Using NLP

- Incident narratives and audit commentary represent a rich source of qualitative risk information. Future extensions could apply natural language processing techniques to extract themes, sentiment, or risk signals from text, integrating these features with structured KRIs to enhance predictive power and early-warning capabilities.

Temporal and Longitudinal Risk Modeling

- While this project primarily relies on aggregated time-window features, future research could investigate time-series modeling approaches to capture evolving risk trajectories and escalation patterns. This may include sequence-based models or hybrid approaches that combine statistical trends with machine learning predictions.

Advanced Model Optimization and Ensemble Techniques

- Further performance and robustness improvements may be achieved through more extensive hyperparameter tuning, model ensembling, or stacking approaches. These techniques could help stabilize predictions across varying risk regimes while preserving interpretability through model-agnostic explainability tools.

References

AIM. (n.d.). Cracking the code of model evaluation & hyperparameter tuning: A deep technical exploration [Study material]. AIM – AI & Machine Learning Programme.

AIM. (n.d.). Week 1: Overview of artificial intelligence and machine learning [Unpublished study material]. AI & Machine Learning Programme.

AIM. (n.d.). Week 3: Python for data analytics [Study material]. AI & Machine Learning Programme.

Bank for International Settlements. (2021). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. https://www.bis.org/publ/othp90.pdf

FIS Global. (n.d.). Risks and ethical implications of AI in financial services. https://www.fisglobal.com/insights/risks-and-ethical-implications-of-ai-in-financial-services

IBM. (n.d.). AI bias. IBM Think. https://www.ibm.com/think/topics/ai-bias

Kaggle. (n.d.). Learn Python. https://www.kaggle.com/learn/python

Molnar, C. (2022). Interpretable machine learning (2nd ed.). https://christophm.github.io/interpretable-ml-book/

W3Schools. (n.d.). Python Introduction. W3Schools. https://www.w3schools.com/python/python_intro.asp

Tools

The following Generative AI tools were used to assist in creating the synthetic dataset, coding, editing and rephrasing, and clarification of technical explanations: Copilot by Microsoft, Gemini by Google, and ChatGPT by OpenAI. The thought process, prompt engineering, analytical decisions, modeling decisions, interpretations, and final conclusions were prepared, reviewed, and validated by the author.