**1. General Description of the AI-Powered Credit Approval System**

Our AI system, the Credit Approval AI Model, is a machine learning-based model specifically designed for credit approval decisions. It is developed by our dedicated team of data scientists and AI experts and was last updated in May 2023. The current version of the system is 1.2.0.

The main purpose of our AI system is to predict an applicant’s creditworthiness based on a variety of financial and personal information. It is developed with a focus on transparency, fairness, and compliance with privacy laws, ensuring that the model’s decisions are explainable and unbiased. It interacts with the FICO HELOC dataset, which includes anonymized customer data such as credit history, income levels, employment status, and other relevant financial details.

Our system caters to three distinct user groups: data scientists, loan officers, and bank customers, each with their own specific requirements. Data scientists play a crucial role in evaluating the machine learning model before it is deployed. Their task is to thoroughly assess its performance and reliability. The final decision rests with the loan officer, who relies on the output generated by the model to make informed judgments. Their role involves carefully considering the model’s insights and using them to reach a conclusive decision. Bank customers, on the other hand, seek clarity regarding the outcome of their loan application. They are interested in understanding the factors that influenced the decision, and our system provides them with the necessary information to satisfy their curiosity. By addressing the needs of these three primary user groups, our system ensures effective utilization across different stages of the loan evaluation process.

The system is primarily used by loan officers to assess credit eligibility and determine the credit quantum. It predicts the likelihood of punctual payments over a period of two years using data derived from credit reports. Our system is built on the IBM AIX360 platform and leverages the robustness and accuracy of the XGBoost machine learning algorithm. It also includes several advanced libraries and tools to guarantee reliable and highly accurate results.

The AI system is implemented using Python and makes use of several open-source libraries, including XGBoost for model training, and AI Fairness 360 (aif360) toolkit for ensuring fairness and reducing bias. It also uses explanatory models from the AI Explainability 360 (aix360) toolkit like BooleanRuleCG, LogisticRuleRegression, and Generalized Linear Rule Models (GLRM) for enhancing model explainability. The XAI tools, Protodash and CEM, are used to explain credit approval decisions to loan officers and bank customers. The current software version requirements are Python 3.6 or higher, XGBoost 1.3.3, aix360 0.2.1, and aif360 0.4.0. Updates to these software components should be applied as they become available to ensure optimal performance and security. The released version, aix360 v0.2.1, has dependencies on the following libraries: joblib>=0.11, scikit-learn>=0.21.2, torch, torchvision, cvxpy>=1.1, Image, tensorflow 1.14, h5py<3.0.0, keras 2.3.1, matplotlib, numpy, pandas, scipy>=0.17, xport, scikit-image, requests, xgboost 1.1.0, bleach>=2.1.0, docutils>=0.13.1, Pygments, osqp, lime 0.1.1.37, shap 0.34.0, nyoka 5.2.0, xmltodict 0.12.0, numba, tqdm, kaggle, and otoc @ git+https://github.com/IBM/otoc@main#egg=otoc. Furthermore, it relies on Nvidia’s DGX System for firmware. The system also utilizes CUDA 10.0 and cuDNN 7.6.5 or later. During the training phase, the model underwent an intensive 8-hour training period on 8x NVIDIA A100 GPUs to achieve optimal performance.

The following are the system requirements for the AI-based Credit Approval System:

• Operating System: Compatibility with Windows, Linux, and macOS.

• Processor: A minimum requirement of an Intel Core i5 processor.

• RAM: At least 8 GB for optimal performance.

• Storage: A minimum of 100 GB of free disk space.

• Graphics Card: A dedicated Nvidia graphics card, such as Nvidia A100.

• Internet Connection: A stable connection for seamless API access.

These requirements may be subject to changes depending on the system configuration and the volume of data involved.

The Credit Approval AI Model is a cloud-based solution, which enables seamless integration and interaction with other hardware and software systems. It is designed to work efficiently on standard server-grade hardware with a robust computational capability. The AI system is placed on the market as a Software as a Service (SaaS) product, allowing banks and financial institutions to subscribe to the service and integrate it into their existing credit approval processes.

While our AI system primarily functions as a standalone service, it can also be integrated as a component of other larger financial software systems or platforms. Given the software nature of our AI system, photographs or illustrations of hardware products are not applicable.

The instructions for the use and installation of our Credit Approval AI Model are provided digitally. The system is hosted on the cloud and accessed via APIs, reducing the need for complex installations. Detailed API documentation is provided, outlining how to send requests to the system and interpret the responses. Regular training and support are provided to ensure users can effectively utilize the system in their credit approval processes.

**User Discretion and Overriding the AI System in a SaaS Context**

The Credit Approval AI Model’s capabilities for discretion and override are carefully designed to be both flexible and secure. As clients subscribe to our service, we acknowledge the need to offer a user-friendly experience that still permits manual intervention when needed. Here are the specifics:

**1 Customizable Policies**: Through our cloud-based dashboard, financial institutions can set custom policies that guide how much weight is given to the AI’s recommendations. This ensures that the AI serves as a tool that augments the loan officer’s judgment rather than dictating a final decision.

**2 Override Permissions**: Institutions can configure user permissions to allow certain levels of staff to override the AI’s recommendations. All overrides are logged securely in the cloud for auditing purposes.

**3 Flag for Manual Review**: Our SaaS platform allows loan officers to flag applications for manual review directly within the user interface. These flagged cases can be routed automatically to designated personnel for further evaluation.

**2. Detailed Description of the AI System’s Elements and Development Process**

**Design and Architecture**

The AI system for credit approval is designed around a robust and efficient machine learning model, XGBoost, known for its excellent performance and accuracy. This model forms the foundation of the AI system, supplemented by other algorithms to ensure transparency, fairness, and explainability in its operations. The system also integrates with third-party tools like the AI Fairness 360 (aif360) toolkit, which is essential for bias reduction and fairness assurance.

The architecture of the AI system is modular, allowing each component to perform its role independently while also integrating seamlessly with the others. Key components include the data pre-processing and normalization module, the model training and evaluation module, and the explainability module.

**Data Source and Pre-processing**

The credit approval system uses the FICO HELOC dataset as its primary data source. The dataset consists of anonymized customer data, including credit history, income levels, employment status, and other relevant financial details. The dataset is composed of a diverse range of customers who have requested a credit line in the range of USD 5,000 - 150,000, which is typically offered by US banks as a percentage of home equity.

The HELOC dataset contains 24 predictor variables and one target variable called RiskPerformance. The table below provides information about the predictor variables and the target variable, including their meaning and monotonicity constraint with respect to the probability of bad = 1:

**Meaning**

**Monotonicity Constraint (with respect to probability of bad = 1)**

Age

None

Consolidated version of risk markers

Monotonically Decreasing

Months Since Oldest Trade Open

Monotonically Decreasing

Months Since Most Recent Trade Open

Monotonically Decreasing

Average Months in File

Monotonically Decreasing

Number Satisfactory Trades

Monotonically Decreasing

Number Trades 60+ Ever

Monotonically Decreasing

Number Trades 90+ Ever

Monotonically Decreasing

Percent Trades Never Delinquent

Monotonically Decreasing

Months Since Most Recent Delinquency

Monotonically Decreasing

Max Delinquency/Public Records Last 12 Months

Values 0-7 are monotonically decreasing

Max Delinquency Ever

Values 2-8 are monotonically decreasing

Number of Total Trades (total number of credit accounts)

No constraint

Number of Trades Open in Last12 Months

Monotonically Increasing

Percent Installment Trades

No constraint

Months Since Most Recent Inquiry excluding last 7 days

Monotonically Decreasing

Number of Inquiries Last 6 Months

Monotonically Increasing

Number of Inquiries Last 6 Months excluding last 7 days. Excluding the last 7 days removes inquiries that are likely due to price comparison shopping.

Monotonically Increasing

Net Fraction Revolving Burden. This is revolving balance divided by credit limit

Monotonically Increasing

Net Fraction Installment Burden. This is installment balance divided by original loan amount

Monotonically Increasing

Number Revolving Trades with Balance

No constraint

Number Installment Trades with Balance

No constraint

Number Bank/National Trades with High Utilization Ratio

Monotonically Increasing

Percent Trades with Balance

No constraint

Risk Performance

Target

The target variable, Risk Performance, is a binary variable that indicates whether an applicant was 90 days past due or worse at least once over a period of 24 months from when the credit account was opened. The value “Bad” indicates that an applicant is a bad credit risk, while the value “Good” indicates that they have made their payments without ever being more than 90 days overdue.

We ensure that all data collection and usage comply with GDPR and other relevant privacy laws. Specifically, all personally identifiable information (PII) is removed or obfuscated to ensure customer privacy.

Before the data is used to train the credit approval AI model, we follow a pre-processing phase that involves several steps to ensure data quality, integrity, fairness, and non-discrimination. We describe these steps below:

**Data Cleaning.** We begin by meticulously removing any inconsistencies, errors, or duplicates in the credit approval dataset to ensure data quality and integrity.

**Feature Selection.** We identify and select the most relevant features that contribute to a customer’s creditworthiness using both domain knowledge and feature importance techniques. This process helps to ensure that only the most relevant features are used to train our model.

**Data Normalization.** We normalize the data to ensure that all features have a similar scale. Normalizing the data helps our model to train more effectively and avoid bias towards any specific feature. This step is crucial in preventing bias towards specific features that may have a larger scale than others.

**Data Bias Reduction.** To ensure fairness and non-discrimination in our credit approval system, we use the Reweighing algorithm to reduce bias in the training data. Specifically, we follow the steps below:

**1 Loading and Preparing Data:** We load the initial credit approval dataset and split it into training and testing datasets. We use the testing dataset to assess the accuracy, fairness, and other metrics during the development of the machine learning model.

**2 Computing Fairness Metric on Original Training Dataset:** We use the AI Fairness 360 (aif360) toolkit to compute the fairness metric on the original training dataset. We identify the protected attribute ‘age’ and define privileged and unprivileged values for that attribute. We then use the mean\_difference method on the BinaryLabelDatasetMetric class to compare the percentage of favorable outcomes for each group. A negative value indicates less favorable outcomes for the unprivileged group.

**3 Mitigating Bias by Transforming the Original Dataset:** To mitigate bias in the training dataset, we use the Reweighing algorithm implemented in the Reweighing class in the aif360.algorithms.preprocessing package. This algorithm transforms the dataset to have more equity in positive outcomes on the protected attribute for the privileged and unprivileged groups. We use the fit and transform methods to perform the transformation, producing a newly transformed training dataset.

**4 Computing Fairness Metric on Transformed Dataset:** We use the mean\_difference method on the BinaryLabelDatasetMetric class again to compute the fairness metric on the transformed training dataset. This metric indicates the percentage of favorable outcomes for each group after the dataset has been transformed. We see that the Reweighing algorithm was highly effective in removing bias, as the difference in mean outcomes is now 0.0. Therefore, there is no advantage for the privileged group.

By following these steps and utilizing the aif360 toolkit, we can ensure that our credit approval system is fair and unbiased, reducing the risk of discrimination and promoting ethical and responsible use of AI-based systems.

Tools for Preprocessing: We use Python libraries like Pandas for data cleaning, and Scikit-Learn for data normalization and feature selection. These tools are widely recognized and well-documented, which makes them ideal for our system.

**Model Training and Validation**

The training phase involves fitting our models to the training dataset. The primary predictive model used in our AI system is XGBoost, a gradient boosting framework renowned for its efficiency and performance.

XGBoost works by iteratively combining multiple decision trees to create a strong learner that can predict whether a loan applicant is creditworthy. It employs a novel technique called gradient boosting to improve the performance of decision trees by combining their predictions.

We train the XGBoost model using the training set, tuning its hyperparameters using the validation set to avoid overfitting and underfitting. The model’s performance is evaluated using the subset accuracy metric, as calculated by the sklearn.metrics.accuracy\_score function. Alongside accuracy, we also evaluate our model using Precision, Recall, F1-score, and AUC-ROC. This allows us to understand not just the number of correct classifications, but also how well the model performs in terms of false positives and false negatives, which is crucial for a credit approval system.

The XGBoost model has several benefits that make it ideal for our credit approval system. It’s capable of handling a large number of features, is robust to outliers in the data, and can model complex non-linear relationships. Furthermore, it provides several parameters that can be tuned to optimize the model’s performance.

In addition to its efficiency and performance, XGBoost is also considered state-of-the-art for tabular data (such as the FICO HELOC data). It has consistently outperformed other popular machine learning algorithms in various data science competitions and real-world applications. XGBoost’s ability to handle large datasets with many features and its flexibility to incorporate a variety of objective functions and evaluation metrics make it a popular choice for many high-risk AI-based systems, including our credit approval system.

The XGBoost model is trained on the pre-processed and normalized data, with its performance evaluated using subset accuracy, as calculated by the sklearn.metrics.accuracy\_score function. Our model is trained to minimize the risk of false negatives (i.e., approving a loan for someone who will default) while balancing the risk of false positives (i.e., denying a loan to someone who would have paid it back). This approach ensures that our model is more accurate.

**System Performance and Testing**

After pre-processing, the data is split into three distinct sets: a training set, a validation set, and a testing set. This partitioning enables us to train our models effectively and evaluate their performance objectively:

**1 Training Set:** The majority of the data (around 70%) is used to train our models.

**2 Validation Set:** A portion of the data (around 15%) is set aside to tune the model’s parameters and prevent overfitting.

**3 Testing Set:** The remaining data (approximately 15%) is used to test the model’s performance on unseen data, which gives us a realistic measure of how the model will perform in real-world scenarios.

The model’s performance is evaluated using the testing set, which gives us a realistic measure of how the model will perform in real-world scenarios. The model is regularly monitored and updated to detect and address emerging biases promptly. This monitoring process also enables the system to adapt to changing trends and operate in compliance with the EU AI Act.

**Explainability Features**

A key feature of our AI system is its explainability. We believe it’s essential for data scientists, loan officers, and bank customers to understand the reasoning behind the credit approval decisions made by the system. The system incorporates built-in human oversight to ensure that the outputs are correct. The models are designed to be transparent, so their workings can be easily understood. They are also monitored during deployment, and adjustments can be made as necessary.

To ensure transparency and explainability in our credit approval decisions, we use three different explanatory models: BooleanRuleCG, LogisticRuleRegression, and Generalized Linear Rule Models (GLRM). These models help us understand the reasoning behind the decisions made by our XGBoost model.

The explanatory models are integrated into the main application through a REST API. The API can generate explanations on-demand for any credit decision made by the AI system.

**1 BooleanRuleCG (BRCG):** The BooleanRuleCG (BRCG) algorithm generates a set of simple boolean rules using a genetic algorithm. These rules can explain the decisions made by the predictive model, even for complex non-linear relationships. BRCG is designed to produce either a disjunctive normal form (DNF) or a conjunctive normal form (CNF) rule to predict whether an applicant will repay the loan on time. A DNF rule corresponds to an individual rule in the rule set, where the AND clauses in the DNF correspond to individual rules in the rule set. The algorithm uses column generation to search the space of possible clauses, which is exponential in size. The training accuracy of the BRCG model is 0.71, and the test accuracy is 0.69.

**2 LogisticRuleRegression (LRR):** The LogisticRuleRegression (LRR) algorithm fits a logistic regression model using rule-based features and is capable of generating interpretable rules that can explain the decision-making process. It uses column generation to generate promising candidates from the space of all possible rules, including unbinarized ordinal features in addition to rules. The complexity parameters lambda0 and lambda1 penalize the number of rules included in the model and the number of conditions in each rule. The training accuracy of the LRR model is 0.74, and the test accuracy is 0.72.

**3 Generalized Linear Rule Model (GLRM):** The Generalized Linear Rule Model (GLRM) algorithm extends the rule-based explanatory model by allowing for more general linear relationships. It produces models that are weighted combinations of rules, which give insight into loan repayment predictability. The algorithm also has the option of combining rules with linear terms. The model output predicts the probability of repaying on time (Y=1). The training accuracy of the CEMExplainer algorithm is 0.73, and the test accuracy is 0.72.

The use of multiple explanatory models allows us to cross-verify the explanations and ensure their reliability. This is particularly important in high-stakes contexts like credit approval, where understanding the reasons behind decisions can have significant impacts.

Our system also supports XAI algorithms for explaining credit approval decisions to loan officers and bank customers. The algorithms we use are:

**• ProtodashExplainer:** This XAI tool is for producing explanations for Loan Officers/employees. The ProtodashExplainer algorithm works with an existing predictive model to show how the customer compares to others who have similar profiles and had similar repayment records to the model’s prediction for the current customer. It helps to evaluate and predict the applicant’s risk. The method takes as input a datapoint (or group of datapoints) that we want to explain with respect to instances in a training set belonging to the same feature space. The method then tries to minimize the maximum mean discrepancy (MMD metric) between the datapoints we want to explain and a prespecified number of instances from the training set that it will select.

**• CEMExplainer:** This XAI tool is for Applicants/bank customers. The CEMExplainer algorithm sits on top of an existing predictive model and helps detect both the features that a bank customer could improve and also further detects the features that will increase the likelihood of approval and those that are within reach for the customer. It provides insight into what – if their application was rejected – they can improve in order to increase the likelihood it will be accepted in the future.

In addition to these XAI tools, we use GitHub to track and audit the development of our AI system. Every change is logged and timestamped to ensure transparency and accountability. We also keep third-party tools up to date and ensure their security to maintain the system’s high standards and comply with all relevant regulations.

**Trade-off Considerations**

Designing the system required careful consideration of various trade-offs. For instance, the use of XGBoost allowed us to handle a large number of features and model complex relationships, but it inherently lacks explainability due to its ensemble nature. This challenge was mitigated by incorporating explainable AI tools and models, balancing the need for high performance and transparency.

Regarding data privacy, we had to balance the need for detailed customer information to ensure accurate credit approval decisions with the requirement to maintain customer privacy. To achieve this, all data was anonymized, and strict data governance protocols were put in place.

**Validation and Compliance**

The AI system’s validation involved a rigorous testing process. The testing procedure was based on well-defined metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). These metrics helped us assess the model’s predictive power and robustness.

To ensure robustness, the system was tested under different scenarios, including stress tests and edge cases. We simulated various situations, including economic downturns and sudden changes in individual credit behavior, to ensure the system’s response was reliable and consistent.

Regarding cybersecurity, the system was designed to comply with the highest standards. It was subjected to rigorous security testing procedures, including penetration testing, vulnerability scanning, and secure code reviews. Access to the system is strictly regulated, and any data exchanges are encrypted using state-of-the-art techniques.

The AI system is also built to comply with the EU AI Act’s requirements. It implements transparency and fairness measures, including explainable AI elements, to ensure that the AI system’s decisions can be understood and contested. All processes are logged and auditable to ensure transparency and accountability.

To ensure ongoing compliance with legal and ethical standards, we conduct quarterly audits. These audits are performed by an internal team and reviewed by an independent third-party to evaluate the system’s fairness, accuracy, and overall performance.

In summary, the AI system for credit approval is a well-rounded and robust system. It balances performance and accuracy with transparency, fairness, and explainability. Rigorous testing and validation measures ensure the system’s robustness and compliance with relevant regulations. Regular updates and human oversight measures help the system adapt to changing trends and maintain its high performance.

**3. Detailed Information about the Monitoring, Functioning, and Control of the AI System**

The credit approval AI system is a state-of-the-art, data-driven platform designed with an overarching focus on ethical guidelines, data privacy regulations, and fairness. It utilizes a combination of powerful machine learning models and explainable AI (XAI) algorithms to ensure transparency and accountability in credit approval decisions. However, like any complex system, it has certain capabilities, limitations, and potential risks that need to be understood and managed. This section provides an in-depth explanation of these aspects along with the necessary human oversight measures.

**System Capabilities**

The AI system is capable of processing large amounts of anonymized customer data, specifically the FICO HELOC dataset, to predict the creditworthiness of loan applicants. The primary predictive model used in our AI system is XGBoost, a gradient boosting framework that combines multiple decision trees to create a strong learner. XGBoost is renowned for its robustness to outliers, ability to handle numerous features, and capacity to model complex non-linear relationships.

Furthermore, the system incorporates three different explanatory models: BooleanRuleCG (BRCG), LogisticRuleRegression (LRR), and Generalized Linear Rule Models (GLRM). These models generate interpretable rules that explain the decisions made by the XGBoost model, ensuring transparency in the decision-making process.

In addition, the system uses Protodash and CEM, XAI algorithms that provide explanations of credit approval decisions to loan officers and bank customers, respectively. These tools allow users to understand the system’s decisions and gain insights into how they can improve their creditworthiness.

**System Limitations and Potential Unintended Outcomes**

While the system is robust and highly efficient, it does have certain limitations. For instance, its performance is reliant on the quality and completeness of the data it processes. Incomplete or biased data could potentially lead to less accurate predictions or unintentional discriminatory outcomes.

Moreover, the AI system, like any predictive model, cannot guarantee absolute accuracy. There’s always a risk of false positives (denying a loan to someone who would have repaid it) and false negatives (approving a loan for someone who will default). Efforts are made to minimize these risks, but they cannot be entirely eliminated.

There may also be potential unintended outcomes if the system’s decisions are interpreted without considering the context. For example, the rules generated by the explanatory models should not be seen as absolute truths but rather as indications based on past data.

**Sources of Risks**

The main source of risk in the AI system comes from data bias. Bias in the training data can lead to discriminatory outcomes, where certain groups of people may be unfairly disadvantaged. To mitigate this, the AI system employs a Reweighing algorithm to reduce bias in the training data, ensuring fairness and non-discrimination.

Another risk is over-reliance on the system’s decisions without human oversight, which could lead to errors going unnoticed. To manage this, the AI system is designed with built-in human oversight to ensure the correctness of the outputs.

**Input Data**

The AI system uses the FICO HELOC dataset, which includes anonymized customer data such as credit history, income levels, employment status, and other relevant financial details. All data used complies with GDPR and other data privacy regulations to ensure the protection of customer information. In addition, all data is thoroughly cleansed, normalized, and processed before being used to train the machine learning models. This ensures the data’s quality and relevance, which is crucial for the accuracy of the system’s predictions.

As appropriate, the AI system also incorporates external datasets and information to refine its predictions. For instance, macroeconomic indicators, like the unemployment rate or inflation rate, can be considered as they can influence an individual’s ability to repay loans. However, the inclusion of such data is contingent on its availability and relevance, and any additional data inputs are carefully vetted for compliance with data privacy laws and regulations.

It’s important to note that, while the AI system can process and learn from a wide range of data, it is designed to avoid sensitive personal attributes (such as race, gender, etc.) in its decision-making process. This is to ensure that the system’s credit approval decisions are fair, unbiased, and non-discriminatory.

In conclusion, the credit approval AI system is a sophisticated tool designed to automate and streamline the credit approval process. It leverages advanced machine learning and explainable AI techniques to ensure accurate, fair, and transparent decisions. However, it is not without its limitations and potential risks, which are carefully managed through comprehensive data processing, algorithmic bias mitigation, and robust human oversight measures. As the system continues to evolve and improve, it promises to be a valuable asset in the credit industry, fostering efficiency, transparency, and fairness.

**Human Oversight Measures Assessment as per Article 14**

Article 14 of the EU AI Act specifies requirements for human oversight of AI systems. The objective is to ensure that human beings remain in control of the system’s actions and can intervene or override decisions where necessary. For our credit approval AI system, we have implemented the following mechanisms for human oversight, along with the assessment criteria for each:

**1. Human-in-the-Loop (HITL)**

**Mechanism**: In the credit approval process, critical decisions made by the AI system are reviewed by a loan officer. The loan officer has the authority to overturn a decision if they find it to be erroneous or unjust.

**Assessment**: The effectiveness of this mechanism is assessed by tracking the frequency and nature of overrides made by loan officers. Metrics such as the “Override Rate” and “Time-to-Override” are monitored. Feedback from loan officers about the system’s decisions also contributes to the assessment.

**2. Audit Trails**

**Mechanism**: Every decision made by the AI system, along with the underlying rationale as provided by the explainability features, is logged in an immutable audit trail.

**Assessment**: Compliance officers periodically review the audit trails to ensure that all actions can be accounted for and can be traced back to a specific point in the decision-making process. The review includes checking the integrity and completeness of the log records.

**3. Periodic Model Review**

**Mechanism**: Data scientists conduct a periodic review of the model to ensure it is performing as expected and not deviating from ethical or legal guidelines.

**Assessment**: The model’s performance metrics, along with fairness and bias metrics, are evaluated during these reviews. Any deviations trigger an in-depth investigation and, if necessary, model retraining.

**4. User Feedback Loop**

**Mechanism**: Customers who undergo the credit approval process are encouraged to provide feedback, especially if they believe a decision was unfair or incorrect.

**Assessment**: Customer feedback is quantitatively and qualitatively analyzed. Metrics such as “Customer Satisfaction Score” (CSAT) and “Net Promoter Score” (NPS) are considered, along with narrative feedback for subjective analysis.

**5. Emergency Off-Switch**

**Mechanism**: The system is equipped with an “emergency off-switch” that can immediately halt the AI system’s operations in case of malfunction or unexpected behavior.

**Assessment**: The responsiveness and effectiveness of the off-switch are tested under simulated emergency conditions. The assessment also includes the time taken to engage the off-switch and the system’s ability to safely and quickly transition control back to human operators.

**Assessment Schedule and Reporting**

Assessment of these human oversight measures is conducted on a bi-annual basis, and findings are reported to both internal stakeholders and relevant regulatory bodies. The reporting includes detailed analyses, key performance indicators, and recommended actions for improvement.

By adhering to Article 14 of the EU AI Act, we aim to ensure that our AI system for credit approval not only meets regulatory requirements but also stands up to ethical scrutiny. This multi-faceted approach to human oversight ensures that our system balances automation with human values and judgment, thereby enhancing trust and reliability.

**4. Detailed Description of the Risk Management System**

In line with Article 9 of the EU AI Act, our credit approval system operates within a robust, continuously evolving risk management system. This system is executed throughout the entire lifecycle of the high-risk AI system, consistently updated, and thoroughly documented to ensure transparency and compliance.

**Risk Identification and Analysis**

To begin with, our system identifies and analyzes known and foreseeable risks associated with its operation. This is achieved by a comprehensive examination of the AI system’s design, the features it uses, and the decision-making processes it follows. Special attention is paid to the potential misuse scenarios and their associated risks. Factors such as data quality, bias in the training data, overfitting, and potential misinterpretation of model explanations are taken into account during risk identification.

**Risk Estimation and Evaluation**

Risks emerging from the intended use of the AI system and under conditions of reasonably foreseeable misuse are estimated and evaluated. Our system uses the FICO Home Equity Line of Credit (HELOC) dataset to model a diverse range of customer scenarios and test the system’s responses. Potential risks such as discriminatory decisions, invasion of privacy, and miscommunication of credit approval decisions are evaluated and quantified.

**Post-Market Monitoring Data Analysis**

In compliance with Article 61, we gather and analyze data from a post-market monitoring system. This continuous process allows us to identify emerging risks, model drift, and unexpected system behavior. By observing the performance and impact of the system in the real world, we can ensure that it remains aligned with its intended purpose and does not create unintended harm.

**Risk Management Measures**

Upon identification and evaluation of risks, appropriate risk management measures are adopted. The system design, including the XGBoost model and the explanatory models, is continuously reviewed and updated to eliminate or reduce risks. Our risk management measures consider the combined application of all system requirements and the state of the art in AI technologies.

Mitigation and control measures are implemented for risks that cannot be eliminated entirely. These measures include regular bias audits, data quality checks, and system performance evaluations. Furthermore, we provide adequate information about the system’s operation and risks to all stakeholders, ensuring transparency.

**Residual Risk Management**

Residual risks associated with each hazard and the overall residual risk of the system are deemed acceptable, provided the system is used for its intended purpose. Residual risks are communicated to users, ensuring they have a comprehensive understanding of the system’s limitations and potential issues.

**Testing**

Our high-risk AI system is rigorously tested to identify the most appropriate risk management measures. The system’s performance is evaluated against defined metrics and probabilistic thresholds appropriate to its intended purpose. Testing procedures are designed to ensure that the AI system performs consistently and in compliance with the requirements set out in this chapter. These procedures are suitable to achieve the system’s purpose and do not go beyond what is necessary to achieve that purpose.

Testing is performed throughout the development process, prior to market placement, and continues after deployment to ensure continual compliance and performance.

**User Considerations**

When developing and updating the risk management system, we give due consideration to the technical knowledge, experience, education, and training expected from the user. We consider the environment in which the system is intended to be used and the potential impact on children.

**Compliance with Directive 2013/36/EU**

As a credit institution, we comply with Directive 2013/36/EU. The aspects described in this document form part of the risk management procedures established by our institution pursuant to Article 74 of the Directive.

**5. A Description of Any Change Made to the System Through Its Lifecycle**

**Version History**

The Credit Approval AI Model has undergone several updates to improve its performance, security, and user experience. Below is a summary of the major changes:

**Version 1.0.0**

• Initial release with XGBoost algorithm for credit approval prediction.

• Basic explainability features using Protodash and CEM.

**Version 1.1.0**

• Introduced LogisticRuleRegression (LRR) and BooleanRuleCG (BRCG) algorithms for enhanced explainability.

• Improved data pre-processing techniques for better data quality.

**Version 1.2.0**

• Added support for additional operating systems, including Linux and macOS.

• Implemented advanced bias detection and mitigation algorithms.

**Rollback Procedures**

In case of critical issues or system failures, rollback procedures are in place to revert the system to its last stable version. All changes are thoroughly tested in a staging environment before deployment to minimize the need for rollbacks.

**6. List of the Harmonised Standards Applied**

In the development and deployment of our credit approval system, we adhered to several harmonised standards to meet the legal and technical requirements of various jurisdictions, including the European Union. These standards encompassed data protection, machine learning, and explainability.

**1 ISO/IEC 27001: Information Security Management:** As we handle sensitive financial data, we adhere to the principles outlined in ISO/IEC 27001, ensuring that robust information security management systems are in place to protect the data we process.

**2 ISO/IEC 27701: Privacy Information Management:** To ensure that we comply with GDPR and other relevant privacy laws, we have adopted ISO/IEC 27701, which provides guidelines for a Privacy Information Management System (PIMS). This helps in managing and reducing privacy risks.

**3 ISO/IEC 38505-1: Governance of Data:** This standard provides guidance on the responsible management of data, particularly as it relates to decision-making processes. Following this standard allows us to ensure the integrity and quality of the data that we use to train our model.

**4 IEEE P7003: Algorithmic Bias Considerations:** To ensure fairness and non-discrimination in our credit approval system, we follow the guidelines set out by the IEEE P7003 standard. This standard provides specific methodologies for addressing bias in algorithms and models. We’ve implemented bias reduction using the Reweighing algorithm as part of our adherence to this standard.

**5 ISO/IEC TR 24028:2020: Artificial Intelligence (AI) — Overview of trustworthiness in Artificial Intelligence:** This standard provides guidance on trustworthiness aspects of AI systems, including robustness, accuracy, privacy, transparency, and explainability. We ensure that our credit approval system adheres to these principles.

In cases where specific harmonised standards have not been applied, we have relied on best practices within the machine learning and AI industry to meet the requirements set out in Title III, Chapter 2. These include:

**• Data Preprocessing and Model Training:** We employ advanced data preprocessing techniques to ensure data quality, integrity, fairness, and non-discrimination. These steps include data cleaning, feature selection, data normalization, and data bias reduction. The primary predictive model used in our system is XGBoost, a gradient boosting framework renowned for its efficiency and performance.

**• Explainability and Transparency:** A key feature of our AI system is its explainability. We believe it’s essential for data scientists, loan officers, and bank customers to understand the reasoning behind the credit approval decisions made by the system. We utilize three different explanatory models: BooleanRuleCG, LogisticRuleRegression, and Generalized Linear Rule Models (GLRM), to ensure transparency and explainability in our credit approval decisions.

**• Use of XAI algorithms:** Our system also supports XAI algorithms for explaining credit approval decisions to loan officers and bank customers. The algorithms we use include Protodash and CEM, providing clear insights into the credit approval decisions.

By adhering to these standards and practices, we ensure that our credit approval system is fair, accountable, transparent, and secure, aligning with the principles and requirements set out in Title III, Chapter 2.

**7. EU Declaration of Conformity**

1 AI system name and type and any additional unambiguous reference allowing identification and traceability of the AI system:

• AI system name: Credit Approval AI Model

• AI system type: Machine learning-based credit approval system

• Additional reference: Version 1.2.0

2 Name and address of the provider or, where applicable, their authorized representative:

• Name: Jane Doe

• Address: 221B Baker Street, London

3 A statement that the EU declaration of conformity is issued under the sole responsibility of the provider:

• We hereby declare that this EU declaration of conformity is issued under the sole responsibility of Jane Doe.

4 A statement that the AI system in question is in conformity with this Regulation and, if applicable, with any other relevant Union legislation that provides for the issuing of an EU declaration of conformity:

• We confirm that the Credit Approval AI Model is in conformity with the AI Act and, where applicable, with any other relevant Union legislation that provides for the issuing of an EU declaration of conformity.

5 References to any relevant harmonized standards used or any other common specification in relation to which conformity is declared:

• The Credit Approval AI Model conforms to the following standards and specifications:

◦ ISO/IEC 27001: Information Security Management

◦ ISO/IEC 27701: Privacy Information Management

◦ ISO/IEC 38505-1: Governance of Data

◦ IEEE P7003: Algorithmic Bias Considerations

◦ ISO/IEC TR 24028:2020: Artificial Intelligence (AI) — Overview of trustworthiness in Artificial Intelligence

6 Where applicable, the name and identification number of the notified body, a description of the conformity assessment procedure performed, and identification of the certificate issued:

• Notified body: English Certification Agency

• Identification number: ECA-6272LON

• Description of conformity assessment procedure: The Credit Approval AI Model underwent a comprehensive evaluation by the English Certification Agency. The assessment included a review of the system’s design, development processes, and adherence to the applicable standards and regulations. Additionally, the AI model was subjected to rigorous testing to ensure its performance, reliability, and compliance with the specified requirements.

• Certificate issued: Certificate of Conformity (Certificate No. AI-5678)

7 Place and date of issue of the declaration, name and function of the person who signed it, as well as an indication for, and on behalf of whom, that person signed:

• Place of issue: London

• Date of issue: May 15, 2023

• Person who signed: Jane Doe

• Function: AI Lead Engineer and Legal Representative

• Signed on behalf of: The ACME Company

**8. Detailed Description of the System in Place to Evaluate the AI System Performance in the Post-market Phase**

After the deployment of the AI system, it is critical to continuously monitor and evaluate its performance to ensure that it maintains its intended functionality, accuracy, and fairness. In the post-market phase, the AI system’s performance evaluation consists of several critical components, including ongoing monitoring, routine performance evaluations, bias detection and mitigation, and system updates. This chapter provides a comprehensive description of the system and procedures we have put in place to effectively evaluate the AI system’s performance in the post-market phase.

**Ongoing Monitoring**

In the post-market phase, the AI system is continuously monitored to track its performance and detect any drifts in the system’s behavior or the data it processes. This ongoing monitoring process involves collecting and analyzing data on the system’s performance metrics, such as accuracy, fairness, and bias. The monitoring process is automated, with real-time alerts set up to notify the team of any significant changes in these metrics that could indicate a problem with the system’s functionality or performance.

The AI system’s performance metrics are compared against predefined thresholds and historical benchmarks to detect any anomalies or unexpected changes. For instance, if the system’s accuracy or fairness metrics fall below the set thresholds, this would trigger an alert for further investigation. The monitoring process also includes tracking the system’s performance across different demographic groups to detect and address any potential bias or discrimination.

**Routine Performance Evaluations**

Besides ongoing monitoring, we also conduct routine performance evaluations of the AI system. These evaluations involve a comprehensive review of the system’s performance metrics and a detailed analysis of the system’s decisions and their impacts.

The routine performance evaluations are conducted on a quarterly basis, or as needed based on the ongoing monitoring results. The evaluations include an analysis of the system’s accuracy and fairness metrics, a review of the system’s credit approval decisions and their outcomes, and a comparison of the system’s performance against other benchmark models or industry standards.

**Bias Detection and Mitigation**

Bias detection and mitigation is a crucial aspect of our post-market evaluation process. We use advanced tools and techniques, such as the AI Fairness 360 (aif360) toolkit, to detect and quantify bias in the AI system’s decisions. The toolkit allows us to compute fairness metrics and compare the percentage of favorable outcomes for different groups.

We also employ a Reweighing algorithm to mitigate any detected bias. This algorithm transforms the training data to ensure equity in positive outcomes across different groups.

If any bias is detected, we conduct a thorough investigation to identify the root cause of the bias and implement appropriate mitigation measures. This could involve retraining the AI model, adjusting the model’s parameters, or revising the data pre-processing procedures.

**System Updates**

Based on the results of the ongoing monitoring and routine performance evaluations, we may implement updates to the AI system to improve its performance or functionality. These updates could involve adjusting the AI model’s parameters, incorporating new features or data sources, or upgrading the AI algorithms.

Each update goes through a rigorous validation and testing process before it is deployed to ensure that it improves the system’s performance and does not introduce new risks or biases. All updates are logged and documented to maintain a transparent record of the system’s evolution and improvements.

**Post-market Audit and Compliance**

To ensure transparency and accountability, we conduct regular audits of our post-market evaluation procedures. These audits involve a thorough review of our monitoring and evaluation practices, performance metrics, bias mitigation measures, and system updates.

We also ensure that our post-market evaluation procedures comply with all relevant regulations, including the GDPR and the EU AI Act. This involves conducting regular compliance checks and maintaining comprehensive documentation of our evaluation procedures and their outcomes. Any non-compliance issues are promptly addressed and corrective actions are taken to bring the procedures back into compliance.

**User Feedback Mechanism**

User feedback is an integral part of our post-market evaluation system. It provides us with valuable insights into the AI system’s real-world performance and its impact on the users. We have implemented a user-friendly feedback mechanism that allows users to report any issues or concerns with the AI system’s decisions.

We actively encourage users to provide feedback on the system’s decisions, its usability, and its impact on their credit approval process. The feedback is regularly reviewed and analyzed to identify any trends or recurring issues that need to be addressed. User feedback is also considered during the routine performance evaluations and system updates to ensure that the AI system meets the users’ needs and expectations.

**Incident Response Plan**

Despite the rigorous monitoring and evaluation procedures, there might be instances where the AI system could encounter unforeseen issues or errors. To handle such incidents effectively and minimize their impact, we have developed an incident response plan.

The plan outlines the procedures to be followed in the event of an incident, including identifying the issue, isolating the affected components of the system, investigating the root cause of the incident, implementing corrective actions, and restoring the system’s functionality. The incident response plan also includes procedures for communicating with users and stakeholders during an incident and for conducting a post-incident review to learn from the incident and prevent similar issues in the future.

In conclusion, our post-market evaluation system is designed to ensure that the AI system maintains its intended functionality and performance, adheres to the principles of fairness and non-discrimination, and continues to meet the needs and expectations of its users. Through ongoing monitoring, routine performance evaluations, bias detection and mitigation, system updates, audits and compliance checks, user feedback, and an effective incident response plan, we continuously evaluate and improve the AI system’s performance in the post-market phase.