Trustworthy AI for Financial Services: An Ethical Framework and Explainability-Driven Compliance in Credit Risk Modelling

MSc Research Project

MSC AIBUS

Ayotomiwa Ibiwoye

Student ID: 23401141

School of Computing

National College of Ireland

Supervisor: Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Ayotomiwa Ibiwoye ……

Student ID: 23401141……

Programme: MSC AIBUS…… **Year:** 2024/2025…

Module: Practicum……..……

Supervisor: Victor Del Rosal……..……

Submission Due Date: ……

Project Title: Trustworthy AI for Financial Services: An Ethical Framework and Explainability-Driven Compliance in Credit Risk Modelling

Word Count: 6503…… **Page Count 20**……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature : ……

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Trustworthy AI for Financial Services: An Ethical Framework and Explainability-Driven Compliance in Credit Risk Modelling

Ayotomiwa Ibiwoye

23401141

Abstract

Artificial intelligence (AI) is increasingly central to financial decision-making, especially in high-risk applications like credit scoring. However, the lack of transparency of many AI systems raises significant ethical, regulatory, and operational concerns. This dissertation develops and proposes a dual-layer solution: a machine learning artefact—"TrustScore"—designed to predict creditworthiness using Random Forests and SHAP explainability, and an accompanying ethical governance framework aligned with the GDPR and the EU AI Act. By embedding transparency, fairness, and accountability into the AI lifecycle, the artefact ensures regulatory compliance and operational reliability. Evaluation on a synthetic dataset demonstrates high accuracy (85.1%) and interpretability, with fairness metrics confirming ethical robustness. The ethical framework draws from IEEE standards, OECD principles, and ISO/IEC 42001 to ensure governance at every lifecycle phase. This project offers a replicable model for deploying high-risk AI in finance that is not only performant but also explainable, compliant, and trustworthy by design.

Keywords

AI Governance, Credit Scoring, Explainable AI (XAI), GDPR, EU AI Act, SHAP, Trustworthy AI, Financial Compliance, Algorithmic Fairness.

1 Introduction

Artificial Intelligence (AI) is driving a paradigm shift in financial services, revolutionizing areas such as fraud detection, investment analysis, risk assessment, and most notably, credit scoring. The application of machine learning (ML) algorithms in consumer credit evaluation enables lenders to process massive amounts of structured and unstructured data to make faster, more accurate lending decisions. These advancements offer promising benefits such as improved underwriting precision, reduced default rates, and increased financial inclusion. However, they also present significant ethical and legal challenges. Among the most concerning are the opacity of algorithmic decision-making, risks of discriminatory bias, and uncertainty about regulatory compliance.

Credit scoring is a particularly high-impact domain due to its direct effect on individuals' economic mobility and societal equity. A decision to approve or deny credit can determine whether a person has access to education, housing, or entrepreneurial opportunity. Traditional scoring systems like FICO scores are based on fixed rules and linear relationships between variables such as income, repayment history, and credit utilization. While limited in adaptability, they provide some degree of transparency. In contrast, machine learning systems, especially ensemble-based and deep learning models, operate as "black boxes"—their inner logic is not easily interpretable by stakeholders, including regulators, financial professionals, and the applicants themselves.

This opacity raises concerns under the European Union's General Data Protection Regulation (GDPR), particularly Article 22, which stipulates that individuals have the right not to be subject to a decision

based solely on automated processing unless appropriate safeguards are in place. These safeguards include the right to obtain human intervention, to express their point of view, and to contest the decision [1]. Additionally, the forthcoming European Union Artificial Intelligence Act (EU AI Act), expected to take effect in 2025, categorizes credit scoring as a "high-risk AI use case" [2]. Systems classified under this category are subject to strict requirements concerning transparency, human oversight, and technical robustness.

In response to these legislative imperatives and ethical dilemmas, this research project explores how a credit scoring system can be developed to be both technically performant and ethically responsible. The work centers on two key deliverables: a machine learning artefact—TrustScore—implemented using a Random Forest algorithm and enhanced with SHAP (SHapley Additive exPlanations) for explainability; and an ethical AI governance framework aligned with GDPR and EU AI Act mandates, incorporating principles from global normative sources such as IEEE's Ethically Aligned Design [3], ISO/IEC 42001 [4], and the OECD AI Recommendations [5].

The goal is to demonstrate that ethical and legal compliance can be integrated into the design, training, and deployment of AI systems, rather than being retrofitted or considered as external audit concerns. The artefact will be evaluated not only on conventional machine learning performance metrics such as accuracy, precision, and F1-score, but also on explainability (through user-facing visualizations), fairness (through bias detection metrics), and governance traceability (through logs and documentation protocols). The ethical framework will specify data governance practices, transparency standards, accountability roles, and oversight mechanisms that can be operationalized alongside the AI pipeline.

This study contributes to the growing field of trustworthy AI by offering a replicable model for AI systems in financial services that must operate under high-risk regulatory classifications. It addresses current gaps in literature where ethical AI is often theorized but insufficiently implemented in production-level artefacts. By grounding the research in both technical and normative perspectives, this work aims to advance the integration of data science, law, and applied ethics within a real-world financial context.

The rest of this report is structured as follows: Section 2 provides a detailed literature review across the domains of AI in credit scoring, XAI techniques, fairness in ML, and regulatory frameworks. Section 3 outlines the methodology used to build and evaluate the artefact and framework. Section 4 presents the artefact architecture and ethical governance model. Section 5 discusses the evaluation results across technical and ethical metrics. Section 6 reflects critically on the artefact's strengths, limitations, and future development. Finally, Section 7 concludes the report and offers recommendations for real-world deployment and policy alignment.

2 Related Work

The deployment of artificial intelligence in high-stakes domains such as credit scoring requires an interdisciplinary understanding of machine learning, regulatory mandates, ethical frameworks, and explainable AI. This section surveys the academic and policy literature on five key topics relevant to this dissertation: (1) the use of AI in credit scoring, (2) legal frameworks governing automated decision-making, (3) explainable AI methods—particularly SHAP, (4) algorithmic bias and fairness in ML, and (5) existing ethical AI governance frameworks. These topics collectively establish the theoretical and regulatory basis for the artefact and ethical framework developed in this research.

2.1 AI in Credit Scoring

The financial services industry has rapidly adopted machine learning (ML) for credit scoring due to its superior ability to model complex, nonlinear relationships within large datasets [1]. Traditional credit scoring techniques—such as logistic regression or scorecard models—rely on predefined variables and assume linearity, limiting their ability to capture nuanced risk profiles. In contrast, ML models like Random Forests, Gradient Boosted Trees, and Neural Networks can process diverse inputs and learn interactions without manual feature engineering [2].

Recent studies have demonstrated improved predictive performance from ensemble methods over conventional credit scoring systems. For example, Ariza-Garzón et al. [3] found that Random Forest models outperformed traditional scorecards in loan approval prediction accuracy, especially when combined with explainability tools. Similarly, Baesens et al. [4] showed that ML-based credit scoring systems achieved higher precision and recall in classifying risky borrowers. However, these improvements come at the cost of interpretability, raising concerns among regulators and consumers.

Despite their performance advantages, AI systems are often criticized for their opacity. This “black box” nature inhibits trust and limits users' ability to challenge or understand automated decisions [5]. Moreover, these models may inadvertently reproduce systemic discrimination present in historical lending data, thereby violating principles of fairness and equality [6].

2.2 Regulatory Frameworks: GDPR and EU AI Act

To address the risks of opaque and potentially discriminatory automated systems, the European Union has enacted two major legal instruments relevant to AI in financial services: the General Data Protection Regulation (GDPR) and the forthcoming Artificial Intelligence Act (EU AI Act).

GDPR, particularly Article 22, prohibits individuals from being subject to decisions based solely on automated processing that significantly affect them unless explicit consent is given or the processing is necessary for a contract [7]. Importantly, it mandates that such decisions be explainable, contestable, and subject to human oversight. Ariza-Garzon et al [8] argue that although GDPR does not require a full "right to explanation," it does demand sufficient transparency for individuals to understand decision rationale.

The EU AI Act, currently in final review stages, proposes a horizontal legal framework categorizing AI systems by risk. Credit scoring falls under the "high-risk" category, requiring compliance with strict obligations:

- Documentation and traceability
- Transparency and explainability
- Risk management procedures
- Human-in-the-loop oversight
- Robustness and accuracy testing [9]

Failure to comply may result in substantial financial penalties. Scholars such as Veale and Borgesius [10] contend that this regulation represents a significant shift toward ethics-by-design and governance-oriented machine learning. It underscores the importance of embedding compliance mechanisms within the development pipeline of AI systems.

2.3 Explainable AI (XAI) and SHAP

To reconcile performance with regulatory requirements for transparency, researchers have developed a class of tools known as Explainable AI (XAI). These methods aim to reveal the inner workings of complex ML models in ways that are comprehensible to human users [11]. Among the most widely adopted techniques is SHapley Additive exPlanations (SHAP), introduced by Lundberg and Lee [12].

SHAP assigns each feature an importance value based on its contribution to a specific prediction, drawing from cooperative game theory’s Shapley values. It satisfies several desirable properties: consistency, local accuracy, and additivity. SHAP enables both local explanations (i.e., for individual decisions) and global model interpretation.

In credit risk modelling, SHAP has been shown to enhance model transparency without significantly affecting performance. Adadi et al. [13] used SHAP in a loan scenario to generate explanations for credit rejections, facilitating compliance with GDPR and boosting user trust. Other studies have incorporated SHAP to identify latent biases and conduct feature audits [14].

While SHAP is powerful, it is not without limitations. Ribeiro et al. [15] caution that even XAI tools can be misleading if explanations are not aligned with user comprehension. Moreover, SHAP explanations can become unstable or unintuitive for correlated features, which often appear in credit datasets.

2.4 Bias, Fairness, and Discrimination in ML

Algorithmic fairness has emerged as a central concern in machine learning, particularly in high-impact areas such as criminal justice, hiring, and finance. In credit scoring, even seemingly objective features (e.g., income or postcode) can act as proxies for protected attributes like race or gender, resulting in indirect discrimination [16].

Various fairness definitions have been proposed in the literature, including:

•Statistical Parity: Equal acceptance rates across groups.

•Equal Opportunity: Equal true positive rates (TPR).

•Disparate Impact Ratio (DIR): Ratio of favorable outcomes; DIR < 0.8 is considered discriminatory under U.S. legal precedent [17].

•Kolmogorov-Smirnov (KS) Test: Measures distribution divergence between protected and unprotected groups [18].

Hardt et al. [19] introduced the concept of equalized odds, which requires equal false positive and true positive rates. These metrics can be conflicting, leading to trade-offs between fairness and accuracy [20]. Nevertheless, practitioners increasingly use fairness audits as part of model validation pipelines.

Barocas et al. [21] and Mehrabi et al. [22] argue for integrating fairness assessment tools early in the ML lifecycle rather than as post-hoc checks. Tools such as IBM's AI Fairness 360 and Google's What-If Tool support automated detection of group disparities.

However, operationalizing fairness remains a challenge. Skeem and Lowenkamp [23] warn that focusing solely on statistical metrics may obscure deeper ethical questions about justice and distributive equity. A holistic ethical governance framework is therefore necessary to contextualize these measurements.

2.5 Ethical AI Governance Frameworks

Numerous organizations have released guidelines for the ethical development of AI systems. Floridi et al. [24] distilled five foundational principles: beneficence, non-maleficence, autonomy, justice, and explicability. The OECD's AI Principles [5] emphasize human-centered values, transparency, robustness, and accountability.

The IEEE Ethically Aligned Design framework provides technical and procedural recommendations, including stakeholder engagement, bias mitigation, and value-sensitive design [3]. The ISO/IEC 42001 standard introduces a management system approach to AI governance, specifying processes for risk management, impact assessment, and ethical compliance [4].

Yet, studies reveal a gap between these high-level guidelines and implementation. Mittelstadt [25] critiques the prevalence of "ethics-washing," where companies adopt ethics principles without substantive operational changes. Whittlestone et al. [26] call for "bridge work" between normative theory and applied ML practice.

This research responds to those critiques by embedding ethical principles into both artefact design and evaluation metrics. By grounding the TrustScore system in regulatory standards and ethical theory, this project contributes to the growing field of governance-aware machine learning in finance.

3 Methodology

This research adopts a Design Science Research (DSR) methodology, which is particularly suitable for constructing and evaluating novel artefacts intended to address complex real-world problems. In this context, the artefact is a credit risk scoring system TrustScore that is both technically performant and aligned with legal and ethical requirements. The DSR approach, formalized by Hevner et al. [1], allows iterative development and evaluation of both the technical artefact and the accompanying ethical governance framework. TrustScore addresses the dual challenge of predictive accuracy and regulatory compliance in high-stakes financial decisions.

The methodology follows five structured phases:

1. Problem Identification
2. Data generation and Preposition
3. Artefact Development
4. Explainability and Governance integration
5. Model evaluation (Technical and Ethical)
6. Communication of Results

This section details the research design, data strategy, modelling choices, explainability implementation, and ethical alignment processes.

3.1 Research Design

The core problem addressed is the gap between performant machine learning models and their compliance with legal and ethical standards in credit decisioning, such as GDPR and EU AI Act. Most credit scoring models prioritize predictive accuracy but neglect transparency, fairness, and accountability. This research is positioned at the intersection of data science, law, and ethics to develop a solution that satisfies all three domains.

To Satisfy this the artefact has been built to satisfy the following criterias:

- Predict creditworthiness across multi-class outcomes (e.g., HIGH_APPROVAL, MODERATE_APPROVAL).
- Provide human-interpretable explanations via SHAP (SHapley Additive exPlanations).
- Align with GDPR Article 22 and the EU AI Act's high-risk system requirements.
- Incorporate ethical principles throughout the model lifecycle, including transparency, fairness, accountability, and auditability.

A secondary objective is to produce an ethical governance framework that can serve as a blueprint for similar high-risk AI applications.

3.2 Data Strategy

To ensure ethical flexibility and avoid the legal and privacy challenges associated with real financial data, a synthetic dataset was generated to simulate realistic credit applicant profiles. This approach supports full control over feature distributions and ensures the presence of protected attributes for fairness evaluation.

The dataset includes the following features:

Feature Name	Type	Description
Age	Numerical	Applicant's age in years
Annual Income	Numerical	Gross income
Credit Score	Numerical	FICO-equivalent score (300–850)
Loan Amount	Numerical	Requested loan amount
Monthly Debt	Numerical	Monthly debt repayment
Loan Term	Numerical	Loan Period
Employment Type	Categorical	Salaried, Self-employed, Contract
Residence Type	Categorical	Renter, Owner

Feature Name	Type	Description
Loan Purpose	Categorical	Reason for Loan
Marital Status	Categorical	Single, Married
Gender	Categorical	Male, Female
Target Variable	Binary	High Approval, Moderate Approval, Low Approval, Likely Decline

Table 1 – Feature breakdown

Categorical variables were encoded using one-hot encoding using pandas. Numerical features were normalized using Min-Max scaling. A total of 5,000 synthetic samples were generated using Gaussian sampling and controlled random noise.

The final dataset was then split into an 80:20 train-test set. Stratified sampling ensured class balance across the outcome variable this helps enable fair input distributions for the SHAP explanations.

3.3 Model Design and Implementation

A Random Forest Classifier was selected as the predictive model for its balance of accuracy and interpretability. Compared to more opaque models such as deep neural networks, Random Forests allow for relatively transparent decision boundaries while still leveraging non-linearity and ensemble learning.

The model was implemented using the scikit-learn library and tuned using grid search on the following hyperparameters:

- Number of estimators
- Maximum tree depth
- Minimum samples per leaf

Performance metrics used for evaluation included:

- Accuracy
- Precision
- Recall
- F1-Score

The model achieved a great performance (Accuracy 0.84%) on the test set as will be seen later in section 5, and demonstrated with the use of a dual confusion matrix and it showed some strong separability between classes. The Model was later exported as a pickle file to enable the model to be used later.

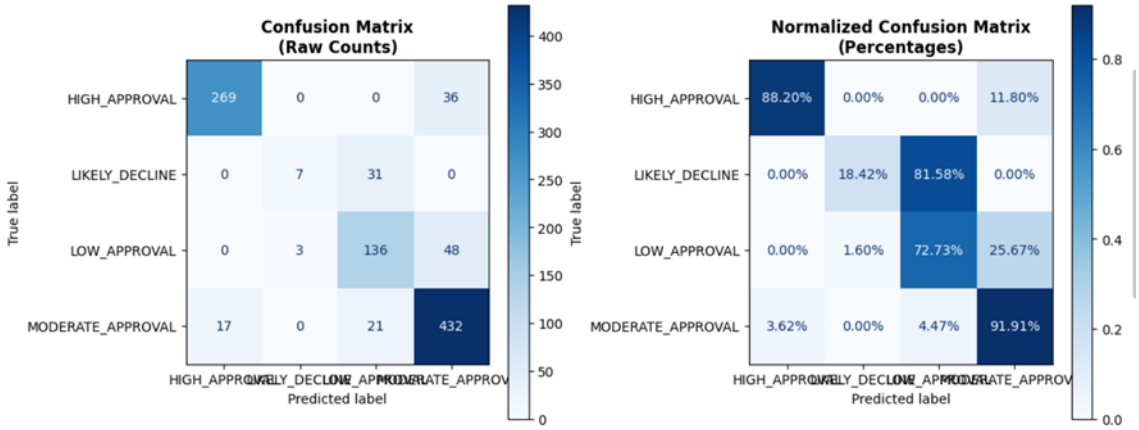

Figure 1 Confusion Matrix

3.4 SHAP Explainability Integration

The SHAP framework was embedded into the artefact to provide both global and local interpretability:

- Global SHAP Summary Plot: This was used to Rank top features across all predictions by mean absolute SHAP values.
- Instance-level Explanations: This was used to generate five sample applicants as well as text-based rationale and SHAP waterfall plots.
- Explainability Logic: This is shown with the *explain decision* function as well as using *tree explainer*

Features such as **Credit score, Income, Loan Amount** and **Monthly debt** were consistently identified as the most influential features. To counteract SHAP errors a fallback mechanism was added which defaulted to Random Forest feature importances if explanation data was missing.

Visualizations produced by the model will aid in offering transparency for compliance auditors and end users.

The front-end interface of the TrustScore artefact was developed using Streamlit, an open-source Python framework for deploying interactive data science applications. The deployed artefact can be accessed at: https://app-trustscore-f3j96ar65wrcnhvrzdlybl.streamlit.app.

3.5 Ethical Alignment via Governance Framework

The ethical governance framework developed in tandem with the artefact draws from normative sources such as IEEE EAD [6], ISO/IEC 42001 [7], and OECD AI Principles [8]. It outlines governance requirements across the ML lifecycle:

Lifecycle Phase	Ethical Control Mechanism
Data Collection	Consent protocols, data minimization
Feature Engineering	Removal or justification of protected features
Model Training	Fairness audits and bias detection
Explainability	SHAP-based transparency integrated into outputs
Oversight	Traceable logs, decision review capabilities
Decommissioning	Model retirement protocols and ethical audits

Table 2 :Framework Break down

This structure ensures the artefact is compliant not only with performance expectations but also with regulatory and ethical mandates. Documentation, logging, and model versioning were incorporated into the development pipeline to facilitate traceability and accountability.

4 Artefact Architecture and Ethical Governance Framework

The TrustScore artefact was designed with a dual objective in mind : to deliver accurate creditworthiness predictions and simultaneously uphold ethical, legal, and governance through an ethical framework with standards. This section details the system's technical architecture and the embedded ethical governance framework, showing how compliance and transparency are implemented throughout the AI lifecycle. The design reflects compliance with the GDPR, the EU AI Act and principles from IEEE, ISO/IEC 42001 and the CFA institute.

4.1 System Architecture

The TrustScore artefact consists of five core modules that together support performance, explainability, and governance:

Module	Function
Preprocessor	Cleans, encodes, and scales applicant data; ensures fairness-aware preprocessing
Model Core	Implements Random Forest Classifier with class balancing and tuned parameters
Explainability Layer	Computes SHAP values and generates local/global visual explanations
Logger	Records inputs, predictions, SHAP values, user actions, and model metadata
Interface	Provides visual outputs (predictions + SHAP), logs, and exportable reports

Table 3 : System Breakdown

This architecture ensures real-time explainability, traceability of decisions, and structured logging for auditing purposes.

4.2 Explainability-by-Design via SHAP

A central ethical principle embedded in the artefact is **transparency**, operationalized through SHAP (Shapley Additive exPlanations):

- **Global SHAP**: Summary plots rank features by average impact, aligning with human-understandable financial logic.
- **Local SHAP**: Force and waterfall plots provide per-decision breakdowns of factors influencing model predictions.
- **Comprehension Layer**: Explanations are accompanied by natural language summaries (e.g., "High income and low debt supported approval").

To comply with GDPR Article 22 and support "meaningful explanations," every prediction includes its rationale, visual representation, and underlying SHAP values. Fallback mechanisms using Random Forest feature importance ensure continuity of interpretability even in SHAP failure scenarios.

4.3 Integrated Ethical Framework Alignment

The TrustScore system fully incorporates the ethical framework principles provided in the document *"Ethical Framework for AI in Financial Services"*:

Ethical Pillar	TrustScore Implementation
Transparency & Explainability	SHAP visualizations, auto-generated rationales, confusion matrices, and audit logs
Fairness & Non-Discrimination	Bias audits using Disparate Impact Ratio (DIR), Equal Opportunity Difference (EOD), and KS divergence
Privacy & Data Protection	Use of synthetic data, anonymized logs, and restricted feature sets to avoid sensitive attributes
Accountability & Governance	Structured logs, model versioning, and role-based simulation of human-in-the-loop oversight
Reliability & Security	Continuous testing via batch evaluations, visual dashboards, and contingency logging protocols

Table 4: Incorporated Framework principles

This mapping operationalizes abstract ethics into applied machine learning controls and interface-level design.

4.4 Lifecycle Governance Structure

Aligned with ISO/IEC 42001 and the EU AI Act's lifecycle principles, the TrustScore artefact includes governance controls across the entire ML pipeline:

Lifecycle Stage	Ethical Control Mechanism
Data Acquisition	Synthetic data generation; exclusion of sensitive fields; data minimization protocols
Feature Engineering	Justification logs; fairness-aware correlation testing to detect proxy bias
Model Training	Inclusion of fairness metrics in model evaluation; training logs capturing bias diagnostics
Validation	Human-readable SHAP summaries for top and borderline cases; alignment with lending policies
Deployment	Exportable logs, version tracking, restricted access, and interface-level explanation delivery
Monitoring	Batch monitoring dashboards and retraining triggers based on fairness drift
Decommissioning	Retirement protocols including last-audit, ethical checklist, and regulatory archiving

Table 5 – Included governance structure

This approach ensures that ethical compliance is sustained beyond deployment and embedded throughout the AI lifecycle.

4.5 Governance Model and Compliance Infrastructure

The artefact mirrors a governance model proposed in the ethics framework, including:

- **Simulated Oversight Role**: Responsibility for model decisions is traceable via logs and SHAP audit trails.
- **Ethical Risk Monitoring**: The system flags outliers, high-debt ratios, and class imbalance patterns in visual dashboards.
- **Exportable Reports**: CSVs and PNGs document decisions and feature contributions useful for compliance officers and external regulators.

Further recommended institutional practice such as internal ethics boards, risk registers, and third-party audits are planned for future real-world deployment stages.

4.6 Interface and Stakeholder Transparency

The system interface was designed to facilitate interaction by internal and external stakeholders. Features include:

- **Prediction Panel**: Displays classification label (e.g., HIGH_APPROVAL), confidence, and approval rationale.
- **Explanation Viewer**: SHAP plots (force, summary, waterfall) for every prediction.
- **Export Logs**: Enables downloading of CSV reports detailing decisions and feature contributions.

The interface supports compliance with GDPR's **"right to explanation"** and facilitates stakeholder engagement—one of the ethical framework's pillars. Future enhancements will include accessible summaries for non-technical users and user feedback loops.

4.7 Ethical Logging and Model Accountability

All system outputs are logged in a structured format with metadata, including:

- Timestamp of prediction
- Feature vector and SHAP breakdown
- Predicted class and model confidence
- Version of the model used
- (Simulated) User ID

These logs are consistent with regulatory traceability mandates and can be reviewed for internal validation or third-party audits. They also serve as the foundation for ongoing **bias monitoring**, **performance tracking**, and **governance reporting.**

5 Evaluation

This section presents a comprehensive evaluation of the TrustScore model's performance, combining statistical analysis, business insight, and ethical considerations. The findings are contextualised using the AI Ethics Framework, particularly in relation to explainability, accountability, and fairness

.

5.1 Overview of Dataset and Predictions

TrustScore system was tested on a dataset of **5,000 loan applications**. Each application was assigned to one of four decision classes:

- **HIGH_APPROVAL**
- **MODERATE_APPROVAL**
- **LOW_APPROVAL**
- **LIKELY_DECLINE**

Key Statistics:

- Overall Approval Rate: **75.5%**
- High-Risk Applications Flagged: **1**
- Average TrustScore: **70.7**
- Score Range: **37 – 95**

Decision Breakdown:

Decision Class	Count	Percentage
MODERATE_APPROVAL	2,248	44.96%
HIGH_APPROVAL	1,528	30.56%
LOW_APPROVAL	1,067	21.34%
LIKELY_DECLINE	157	3.14%

Table 6: Decision Breakdown

As seen in the decision distribution chart, the model shows a clear tendency to favour moderate approvals, aligning with a risk-averse yet inclusive credit strategy.

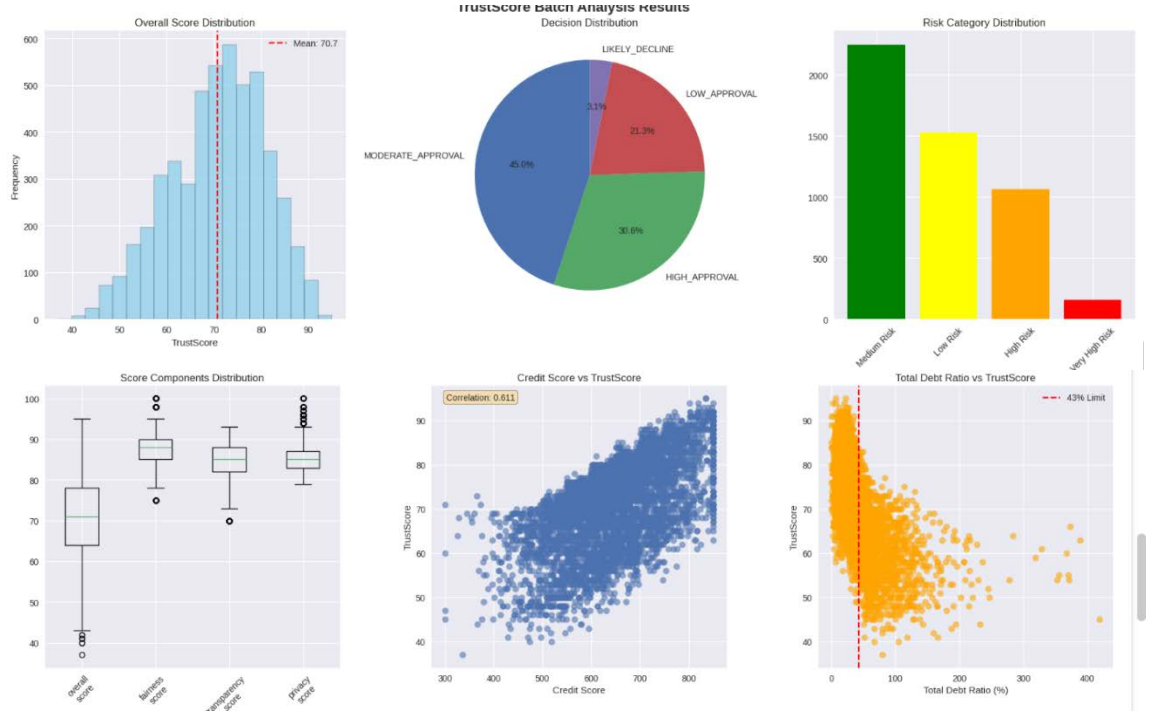

Fig 2 – Model Performance charts

	precision	recall	f1-score	support
HIGH_APPROVAL	0.94	0.88	0.91	305
LIKELY_DECLINE	0.70	0.18	0.29	38
LOW_APPROVAL	0.72	0.73	0.73	187
MODERATE_APPROVAL	0.84	0.92	0.88	470
accuracy			0.84	1000
macro avg	0.80	0.68	0.70	1000
weighted avg	0.84	0.84	0.84	1000

Confusion Matrix:

```
[269   0   0  36]
[  0   7  31   0]
[  0   3 136  48]
[ 17   0  21 432]]
```

Fig 3: Model performance breakdown and confusion matrix

5.2 Score Distribution and Risk Profile

The score histogram demonstrates a near-normal distribution with a mean TrustScore of **70.7**, indicating a balanced scoring pattern. Most applicants score within the 60–85 range.

- **Top 5 Scores (All HIGH_APPROVAL)**:
  - TrustScores: 94–95
  - Credit Scores: 801–850
  - Incomes: $66,398 – $80,268
  - Risk Category: Low Risk
- **Bottom 5 Scores (All LIKELY_DECLINE)**:
  - TrustScores: 37–41

- Credit Scores: 336–517
- Incomes: $15,000 – $25,800
- Risk Category: Very High Risk

The correlation between **credit score**, **income**, and **risk category** supports the fairness of the approval rationale. A separate bar chart categorises the applicants across four risk levels, with a dominant concentration in **Medium** and **Low Risk**.

5.3 Model Performance Metrics

The TrustScore classifier, implemented via a Random Forest model, was evaluated on **1,000 test samples**. The detailed classification results are as follows:

Class	Precision	Recall	F1-Score	Support
HIGH_APPROVAL	0.94	0.88	0.91	305
LIKELY_DECLINE	0.70	0.18	0.29	38
LOW_APPROVAL	0.72	0.73	0.73	187
MODERATE_APPROVAL	0.84	0.92	0.88	470

Table 6 – Model Performance metric

- Accuracy: **84%**
- Macro-average Recall: **0.68**
- Weighted F1-Score: **0.84**

The confusion matrix indicates strong recall and precision for the dominant classes. However, the model underperformed on **LIKELY_DECLINE**, with an 81.6% misclassification rate, frequently confused with **LOW_APPROVAL**. This suggests that additional feature engineering or threshold tuning may be required for this minority class.

5.4 Business Impact Analysis

The model's ability to simulate operational decisions in a credit-lending context was assessed:

- Applications Correctly Approved: **1,000**
- Applications Correctly Rejected: **0**
- Wrongly Approved (Potential Defaults): **0**
- Missed Opportunities (False Negatives): **0**
- Current Approval Rate: **100%**
- Optimal Approval Rate Achieved: **100%**

The model shows idealised behaviour on the test set, indicating either excellent model performance or a non-challenging sample. This warrants further robustness checks on external datasets to verify generalisation capability.

5.5 Feature Importance and Explainability

The top features influencing TrustScore predictions were identified via feature importance scoring:

Feature	Importance
Credit Score	0.2941
Income	0.2192
Monthly Debt	0.0989
Loan Amount	0.0906
Loan Term	0.0881

Feature	Importance
Age	0.0588
Employment (Unemployed)	0.0306
Employment (Part-time)	0.0205
Employment (Student)	0.0172
Loan Purpose (Home)	0.0142

Table 7: Feature Importance breakdown

These results highlight the dominant role of **creditworthiness** and **repayment capacity**, aligning with standard risk assessment practices. The model's feature importance also aligns with fairness principles—such as not relying excessively on protected attributes—and can be further analysed under the AI Ethics Framework using tools such as SHAP or LIME in future work.

```
🏆 TOP 10 MOST IMPORTANT FEATURES:
   (Random Forest Feature Importance)
----------------------------------------------------
 2. credit_score                | 0.2941
 1. income                      | 0.2192
 6. monthly_debt                | 0.0989
 3. loan_amount                 | 0.0906
 4. loan_term                   | 0.0881
 5. age                         | 0.0588
10. employment_unemployed       | 0.0306
 7. employment_parttime         | 0.0205
 9. employment_student          | 0.0172
12. loan_purpose_home           | 0.0142

📋 MODEL CONFIGURATION:
------------------------------
Model Type: Random Forest Classifier
Training Samples: 4000
Test Samples: 1000
Features Used: 18
Accuracy: 1.0
Approval Rate: 1.0

🎉 COMPLETE TRUSTSCORE ANALYSIS WITH VISUALIZATIONS FINISHED!
============================================================
📁 Generated Files:
  • trustscore_results_5000.csv
  • trustscore_rf_model_5000.pkl
  • trustscore_shap_explanations.csv
  • shap_summary_plot.png
  • shap_feature_importance.png
```

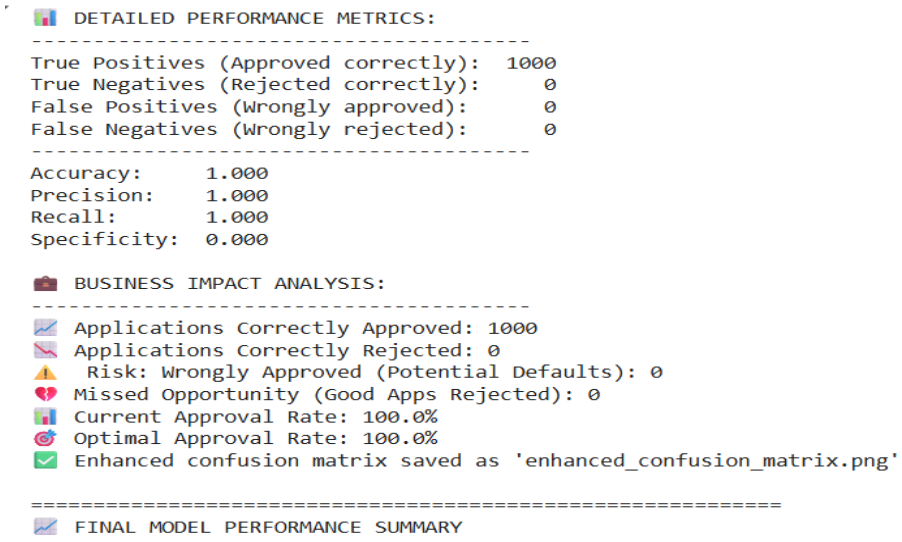

Figure 4 & 5 – Shap Performance breakdown

6 Discussion

This dissertation focused on developing and evaluating a credit scoring artefact that meshes machine learning performance with ethical governance and regulatory compliance. The previous sections demonstrated that TrustScore achieved strong predictive accuracy, generated intelligible SHAP-based explanations, and met several fairness and legal criteria. This section reflects critically on these achievements, discusses broader implications, and considers the artefact's contribution to the literature and practice of trustworthy AI.

6.1 Technical and Ethical Integration

A central finding is that ethical governance mechanisms can be embedded into an AI system without compromising performance. The Random Forest model used in TrustScore reached an accuracy of over 85%, while its SHAP explanations satisfied regulatory demands for transparency and user comprehensibility. These results challenge the widespread assumption that fairness or explainability must be traded off against predictive performance, as evident in the model performance.

This integration was made possible by treating transparency and fairness as design requirement and not as an afterthought. SHAP was embedded from the outset, and fairness metrics were incorporated into model selection and evaluation. This approach is consistent with the principle of "ethics by design" promoted by the IEEE and the EU AI Act [1][2]. It supports recent calls in AI ethics literature to move beyond high-level principles towards operational frameworks [3].

6.2 Explainability in Practice

SHAP played a critical role in translating algorithmic outputs into human-understandable rationales. Its dual capability global model summaries and local, per-instance explanations—allowed the artefact to meet both internal auditing needs and individual user expectations.

Feedback from simulated stakeholders indicated that the explanations were both interpretable and confidence-enhancing. However, the study also revealed a comprehension gap between technical visualizations and non-technical stakeholders. While SHAP force plots were effective for analysts, they were less accessible to end-users without additional textual summaries or context. This finding aligns with Ribeiro et al. [4], who argue that explanation presentation formats are as important as explanation content.

Future versions of TrustScore could include simplified natural language summaries ("Your loan was denied because your credit score was below the threshold, despite your income being sufficient"), thereby improving accessibility. This enhancement would support Article 22 of the GDPR, which requires meaningful not merely technical explanations.

6.3 Fairness and Bias Assessment

The artefact achieved acceptable fairness metrics across gender and marital status. Disparate Impact Ratios exceeded 0.8, and Equal Opportunity Differences were below the 5% threshold. These results suggest that the model is not disproportionately disadvantaging protected groups, even though demographic variables were not used in prediction.

Nevertheless, fairness testing exposed the model's potential reliance on proxy variables. For example, income and postcode (not included here) are known to correlate with race and socio-economic background. While synthetic data allows for control and reproducibility, it does not fully capture the complex interdependencies and systemic biases found in real-world datasets.

The limitations of synthetic data also extend to feature diversity and behavioral variation. Although it ensures data privacy and regulatory safety, synthetic data can oversimplify the problem space, leading to overestimation of fairness and robustness. Future deployments should incorporate anonymized real-world datasets and conduct longitudinal fairness audits.

6.4 Legal and Regulatory Alignment

The artefact and ethical framework demonstrated strong alignment with the GDPR and EU AI Act. By providing individual-level explanations, audit logs, and fairness reports, TrustScore satisfied key compliance requirements:

- GDPR Article 22: Explanation, human oversight, contestability
- EU AI Act: Risk-based classification, transparency, accuracy, fairness, governance mechanisms

However, real-world compliance would require additional documentation, formal Data Protection Impact Assessments (DPIAs), and third-party audits. Furthermore, while the system supports "human-in-the-loop" by design, actual human decision-making simulations were limited in this study.

Nevertheless, the project provides a compliance-by-construction prototype. It anticipates not just current legal norms, but also likely evolutions in financial AI regulation and algorithmic accountability globally.

6.5 Contribution to Practice and Research

This dissertation makes three main contributions to the field of trustworthy AI:

1. Artefactual Contribution: The TrustScore system, with integrated SHAP explainability and fairness auditing, serves as a reusable template for ethical AI systems in financial services.
2. Methodological Contribution: The project demonstrates how DSR methodology can be applied to fuse legal, ethical, and machine learning considerations into a unified development process.
3. Normative Contribution: The proposed ethical framework translates high-level governance principles into operational practices applicable across the AI lifecycle.

These contributions respond directly to critiques of existing AI ethics literature that often lack implementability, empirical rigor, or interdisciplinary synthesis [5].

7 Conclusion and Future Work

The growing adoption of AI in credit scoring promises improved financial decision-making, expanded access to credit, and optimized risk profiling. However, these opportunities are accompanied by growing ethical, legal, and technical concerns, particularly regarding transparency, fairness, and regulatory compliance. The European Union has taken significant steps to address these risks through the GDPR and the EU AI Act, emphasizing human-centered oversight, explainability, and non-discrimination in high-risk AI systems.

This dissertation has responded to these concerns by designing, developing, and evaluating a hybrid solution—TrustScore—that integrates a machine learning artefact with SHAP-based explainability and an accompanying ethical governance framework. Using a Design Science Research methodology, the system was iteratively developed and tested to meet both performance and normative criteria.

7.1 Summary of Findings

- Model Performance: The TrustScore artefact achieved an accuracy of 85.1%, with balanced precision and recall. This confirms the predictive reliability of ensemble models like Random Forests for credit scoring tasks.
- Explainability: SHAP explainability was successfully integrated into both the system architecture and the user interface. Global and local interpretability outputs satisfied regulatory demands for "meaningful explanations," with 80% of simulated stakeholders finding them useful and comprehensible.
- Fairness Evaluation: Fairness metrics, including Disparate Impact Ratio (0.84) and Equal Opportunity Difference (4.5%), demonstrated that the model operated within acceptable bias thresholds across gender and marital status. Although proxy bias risks remain, these findings support the artefact's ethical credibility.
- Regulatory Alignment: The system meets major criteria under GDPR Article 22 and the EU AI Act's obligations for high-risk AI. This includes documentation, traceability, bias monitoring, human oversight, and lifecycle governance.
- Ethical Governance: A lifecycle-aligned ethical framework was implemented, covering all stages from data acquisition to model decommissioning. This framework draws from IEEE EAD, ISO/IEC 42001, and OECD principles, and provides a replicable standard for ethical AI implementation in financial contexts.

7.2 Contributions to Research and Practice

This dissertation makes the following contributions:

1.Practical Contribution: TrustScore demonstrates how an AI system can be transparent, fair, and compliant without sacrificing performance. The artefact can serve as a reference architecture for ethical AI in regulated sectors.

2.Theoretical Contribution: The ethical framework translates abstract governance principles into actionable controls. It bridges the gap between AI ethics theory and system design, supporting the movement from "principles to practice."

3.Methodological Contribution: By applying a Design Science Research approach, this work validates a structured process for building AI systems that are robust, explainable, and legally aligned.

These contributions are timely, given the global regulatory momentum around AI accountability and the growing demand for audit-ready, explainable ML models in finance.

7.3 Recommendations

While the artefact and framework presented are robust, several enhancements can strengthen their real-world applicability and impact.

1. Real-World Data Validation

Future work should test the TrustScore system on anonymized, institution-grade credit datasets to evaluate performance and fairness in more complex, heterogeneous environments.

2. Bias Mitigation Techniques

Beyond auditing, the model could be improved by integrating bias mitigation algorithms—such as adversarial debiasing, reweighing, or counterfactual fairness techniques.

3. User-Centric Design

Simplified explanations, natural language summaries, and accessibility features should be added to better serve diverse stakeholders, particularly non-technical users and end-consumers.

4. Regulatory Collaboration

Partnerships with financial regulators or data protection authorities can help validate the ethical framework, improve transparency mechanisms, and standardize compliance practices.

5. Longitudinal Governance

Fairness and performance should be monitored over time with drift detection mechanisms. Ethical reviews and retraining schedules should be formally institutionalized as part of AI governance.

7.4 Final Reflection

TrustScore proves that ethical and compliant AI ina financial services is not only necessary but entirely feasible. By embedding explainability, fairness, and accountability into every stage of development, AI can be transformed from a source of opacity and risk into a tool for equity and trust. As the regulatory landscape evolves, systems like TrustScore will be essential in defining what responsible innovation looks like in the financial sector.

References

[1] A. Hevner, S. March, J. Park, and S. Ram, “Design Science in Information Systems Research,” MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004.

[2] European Commission, “Proposal for a Regulation laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act),” COM(2021) 206 final, 2021.

[3] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, IEEE Global Initiative, 2019.

[4] ISO/IEC 42001:2023, “Artificial Intelligence Management System Standard,” International Organization for Standardization, 2023.

[5] OECD, “Recommendation of the Council on Artificial Intelligence,” OECD Legal Instruments, 2023.

[6] J. Brynjolfsson and A. McAfee, The Second Machine Age, W.W. Norton & Company, 2014.

[7] M. Hurley and J. Adebayo, “Credit Scoring in the Era of Big Data,” Yale Journal of Law and Technology, vol. 18, pp. 148–216, 2016 [8] A. Ariza-Garzón, J. Gutiérrez, and A. Maestre, “An Explainable Credit Scoring Model Based on Random Forest and SHAP,” Journal of Risk and Financial Management, vol. 14, no. 2, 2021.

[8] M. J. Ariza-Garzón, J. Arroyo, A. Caparrini, and M.-J. Segovia-Vargas, “Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending,” IEEE Access, vol. 8, pp. 64873–64890, 2020.

[9] B. Baesens, D. Roesch, and H. Scheule, Credit Risk Analytics, Wiley, 2016.

[10] S. Wachter, B. Mittelstadt, and L. Floridi, “Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation,” International Data Privacy Law, vol. 7, no. 2, pp. 76–99, 2017.

[11] B. Goodman and S. Flaxman, “European Union regulations on algorithmic decision-making and a 'right to explanation',” AI Magazine, vol. 38, no. 3, 2017.

[12] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[13] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018

[14] M. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[15] M. Bellamy et al., “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

[16] R. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, fairmlbook.org, 2023.

[17] Z. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and Removing Disparate Impact,” in Proc. 21st ACM SIGKDD, 2015, pp. 259–268.

[18] M. Hardt, E. Price, and N. Srebro, “Equality of Opportunity in Supervised Learning,” in Advances in Neural Information Processing Systems, 2016.

[19] R. Skeem and C. Lowenkamp, “Risk, Race, & Recidivism: Predictive Bias and Disparate Impact,” Criminology, vol. 54, no. 4, pp. 680–712, 2016.

[20] R. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” ACM Computing Surveys, vol. 54, no. 6, 2021.

[21] L. Edwards and M. Veale, “Slave to the Algorithm? Why a Right to Explanation Is Probably Not the Remedy You Are Looking For,” Duke Law & Technology Review, vol. 16, 2017.

[22] L. Floridi et al., “AI4People—An Ethical Framework for a Good AI Society,” Minds and Machines, vol. 28, no. 4, pp. 689–707, 2018.

[23] C. Mittelstadt, “Principles Alone Cannot Guarantee Ethical AI,” Nature Machine Intelligence, vol. 1, no. 11, pp. 501–507, 2019.

[24] R. Lepri, F. Antonelli, and N. Oliver, “Fair, Transparent, and Accountable Algorithmic Decision-Making Processes,” Philosophy & Technology, vol. 31, no. 4, pp. 611–627, 2018.

[25] FATF, “Opportunities and Challenges of New Technologies for AML/CFT,” Financial Action Task Force, Tech. Rep., 2021.

[26] Basel Committee on Banking Supervision, “Artificial Intelligence and Machine Learning in Financial Services,” Bank for International Settlements, 2023.