A Deep Dive into Unsupervised, Anomaly-Based Credit Risk Assessment

A Comprehensive Technical Report

Author: AKSHAY S HARITS

A final project report documenting the design, implementation, and evaluation of a credit risk model using alternative financial data.

September 1, 2025

Abstract

This report details the comprehensive journey of developing FinShield, a sophisticated credit risk assessment model designed for underbanked populations. The project tackles the critical challenge of evaluating creditworthiness in the absence of traditional financial data by leveraging alternative data, specifically customer transaction histories. Through an iterative process, the project evolved from initial supervised and clustering-based approaches, which were found to be flawed, to a final, robust unsupervised anomaly detection system.

The final model employs an **Isolation Forest** algorithm to identify users with anomalous financial behaviors. These mathematical deviations are translated into an initial risk score which is then refined by a **logical adjustment engine**. This engine applies discounts to users exhibiting strong positive financial signals, ensuring that ”good outliers” are scored appropriately. A key contribution of this work is its emphasis on transparency; model predictions are fully explained using **SHAP (SHapley Additive exPlanations)**, providing feature-level justifications for each user’s score. This report provides a deep dive into the dataset, the methodological evolution, the advanced feature engineering, and a thorough analysis of the final results, demonstrating a successful solution that is objective, scalable, and transparent.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Challenge of Financial Inclusion

In the global financial landscape, a significant barrier to economic growth for a large segment of the population is the lack of access to credit. Traditional financial institutions heavily rely on historical credit data to assess an individual's creditworthiness. This paradigm inherently excludes ”credit invisible” individuals—those who have limited interaction with the formal banking sector. This creates a cyclical problem: without a credit history, one cannot access credit, and without credit, it is difficult to build a credit history. This project directly confronts this challenge.

1.2 Project Goal and Core Objectives

The primary goal of this project is to design, build, and validate a credit risk assessment model that can effectively serve under-credited populations by using alternative data. The model must be robust, fair, and transparent, aligning with modern principles of responsible AI.

The core objectives were defined as follows:

1. **Leverage Alternative Data:** The model's primary input must be non-traditional data. We focused on raw transaction history as a rich source of behavioral information.

2. **Ensure Objectivity:** The scoring mechanism must be fundamentally data-driven, minimizing subjective human intervention and potential for bias in the final score calculation.

3. **Achieve Transparency:** The model must be explainable, providing a clear, feature-by-feature rationale for any given risk score.

4. **Deliver a Quantifiable Score:** The model's output must be a single, continuous, and intuitive numeric score on a 0-100 scale, where a higher score indicates a higher risk.

5. **Validate on Unseen Data:** The system must be built using a rigorous train/test methodology to prove its ability to generalize to new customers.

1.3 Summary of the Final Architecture

The project's journey led to a purely unsupervised anomaly detection pipeline, enhanced with a logical post-processing layer. The final system operates as follows:

1. It ingests raw transaction and user data.
2. A sophisticated **feature engineering** module constructs a multi-dimensional ”financial fingerprint” for each user.
3. An **Isolation Forest** model calculates a raw **anomaly score**, quantifying how much a user's behavior deviates from the norm.
4. This score is converted to a preliminary 0-100 risk score based on its percentile rank.
5. A **Logical Adjustment Engine** applies discounts to this score based on ”Golden Features” (e.g., being debt-free), ensuring good outliers are not penalized.
6. The final prediction is explained using **SHAP**, revealing the exact features that contributed to the user's raw anomaly score.

Chapter 2

Dataset & Exploratory Analysis

2.1 Dataset Overview

The project utilizes a comprehensive financial dataset from a 2024 AI Hackathon by Caixabank Tech, publicly available on Kaggle. It contains anonymized data from a banking institution throughout the 2010s. This rich, multi-faceted dataset is designed for a variety of analytical tasks, including fraud detection and customer behavior analysis.

2.1.1 Data Sources Utilized

Our model was built by integrating information from three primary data files:

- `transactions_data.csv`: This is the core of our alternative data, containing millions of transaction records. Each record includes a client ID, a transaction amount, a timestamp, and a merchant category code (MCC), which allows us to analyze the volume, velocity, and nature of user spending.

- `users_data.csv`: This file provides the demographic and baseline financial context for each customer. Key fields include age, stated yearly income, total debt, and geographic location (city, state). This information is crucial for creating ratio-based features and understanding the user's background.

- `mcc_codes.json`: A mapping of the numerical Merchant Category Codes to human-readable descriptions (e.g., 5411 to 'Grocery Stores'). This file is essential for the feature engineering step, where we categorize transactions into logical groups like ”Wants,” ”Needs,” and ”Vice.”

2.1.2 Data Dictionary

A summary of the most critical fields used in this project is provided in Table 2.1.

Table 2.1: Key Data Fields

Field Name	**Source File**	**Description**
`client_id`	transactions, users	Unique identifier for each customer.
`amount`	transactions	The monetary value of the transaction.
`date`	transactions	Timestamp of the transaction.
`mcc`	transactions	Merchant Category Code for the transaction.
`current_age`	users	The age of the customer.
`yearly_income`	users	The stated annual income of the customer.
`total_debt`	users	The total outstanding debt of the customer.
`city, state`	users	Geographic information for the customer.

Chapter 3

Methodological Evolution: A Journey in Model Selection

The path to the final model was not linear. It was an iterative process of proposing a solution, discovering its limitations, and using those insights to build a better approach. This chapter documents that journey.

3.1 Attempt 1: The Supervised Learning Dead End

3.1.1 Hypothesis

The initial hypothesis was that we could train a supervised classification model to predict the probability that a user is ”high-risk,” where ”high-risk” was defined as having one or more fraudulent transactions as per `train_fraud_labels.json`.

3.1.2 Proposed Model: XGBoost

The proposed algorithm was XGBoost (Extreme Gradient Boosting), a powerful and widely used tree-based ensemble method. It is known for its high performance on structured, tabular data like our feature set.

3.1.3 Failure Analysis

As discovered during exploratory analysis, the target variable itself was fundamentally flawed. With 98% of the active user base labeled as ”high-risk,” the model would face a severe class imbalance, but more importantly, the majority class was meaningless. The model’s optimal strategy would be to simply predict ”high-risk” for nearly every user, resulting in high accuracy but zero practical utility for credit assessment. This approach failed because the provided labels did not align with a meaningful definition of credit risk.

3.2 Attempt 2: The Unsupervised Clustering Diversion

3.2.1 New Hypothesis

Since the labels were unreliable, the next hypothesis was that we could uncover natural groupings of users based on their financial behavior using unsupervised clustering. These clusters, or ”personas,” could then be analyzed to determine their inherent risk level.

3.2.2 Proposed Model: Gaussian Mixture Models (GMM)

GMM was chosen over simpler algorithms like K-Means because of its flexibility. GMM is a probabilistic model that assumes the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. This allows it to model clusters that are not necessarily spherical and provides a probability of assignment for each point to each cluster.

3.2.3 The Subjectivity Trap

This approach successfully identified distinct personas (e.g., ”High-Volume Spenders,” ”Frequent & Frugal Shoppers”). However, it introduced a critical new problem: **subjectivity**. To translate a persona into a 0-100 score, a human analyst had to:

1. Interpret the meaning of each cluster’s average features.
2. Subjectively label each cluster with a risk category (e.g., ”Low,” ”Medium,” ”High”).
3. Manually define a scoring range for each category.

This process was not truly data-driven and was susceptible to human bias, violating the core project objective of objectivity..

3.3 The Anomaly Detection Breakthrough

The final hypothesis reframed the problem from classification to deviation:

How much does an individual’s financial behavior deviate from the stable, predictable norm?

3.3.1 Why Unsupervised Anomaly Detection is Ideal

This approach is uniquely suited for the ”credit invisible” population for several profound reasons:

It Thrives in the Absence of Labels. By definition, underbanked populations lack the historical ”default” vs. ”non-default” labels required for supervised learning. Anomaly detection does not need them.

It is Fundamentally Data-Driven and Objective. The model learns the definition of ”normal” directly from the data. This minimizes the human bias that can creep into supervised models trained on historical data, which may reflect past discriminatory lending practices.

The Core Assumption: Normalcy is a Proxy for Low Risk. The underlying hypothesis is that the majority of people’s collective behavior forms a stable, predictable, and low-risk pattern. A user whose financial fingerprint is highly unusual is, by definition, a higher risk.

Analogy: Spotting a Counterfeit Bill. To train an expert to spot counterfeit currency, you don’t show them every fake bill ever made. You train them to know every detail of a *real* bill. Once they deeply understand ”normal,” they can instantly spot anything that deviates. Our model works the same way.

Chapter 4

The Data & Feature Engineering Pipeline

4.1 Advanced Feature Engineering

The feature engineering process was refined to create a lean but powerful set of predictors, focusing on advanced stability metrics, standard financial ratios, and semantic signals of behavior.

Table 4.1: Final Engineered Behavioral Features

Feature Name	**Mathematical Definition**
`spending_cv`	$\frac{\sigma_{\text{monthly spend}}}{\mu_{\text{monthly spend}}}$
`dti_ratio`	$\frac{\text{Total Debt}}{\text{Yearly Income}}$
`high_risk_behavior_flag`	Binary flag for distress or high gambling txns
`financial_discipline_score`	$1 - \text{Avg}\Big(\text{Rank}(\texttt{spending\_cv}), \text{Rank}(\texttt{wants\_ratio})\Big)$
`age_cushion_interaction`	$\frac{\text{Age Risk Score}}{\text{Financial Cushion Score} + 1}$
`savings_x_discipline`	$\texttt{savings\_rate} \times \texttt{discipline\_score}$

Spending Coefficient of Variation (`spending_cv`). This feature measures the stability of a user's spending behavior by comparing variability to the mean. A lower coefficient indicates disciplined, predictable spending, which is a strong positive credit signal.

Debt-to-Income Ratio (`dti_ratio`). This classical financial feature captures leverage by dividing total debt by yearly income. A higher ratio suggests financial stress and is strongly associated with default risk.

High-Risk Behavior Flag (`high_risk_behavior_flag`). A binary indicator triggered by transactions in categories such as payday loans, pawn shops, or excessive gambling. Even a single occurrence is considered a red-flag for financial distress.

Financial Discipline Score (`financial_discipline_score`). A composite engineered feature that combines spending stability and the ratio of discretionary ("wants") to essential ("needs") expenses. Higher values correspond to responsible and well-regulated financial behavior.

Age-Cushion Interaction (`age_cushion_interaction`). An interaction term that models how age-related risk compounds with lack of financial cushion (savings or low leverage). Younger users with weak cushions are less risky than older users in the same position.

Savings × Discipline (`savings_x_discipline`). A positive interaction term rewarding individuals who simultaneously maintain high savings rates and demonstrate disciplined spending behavior.

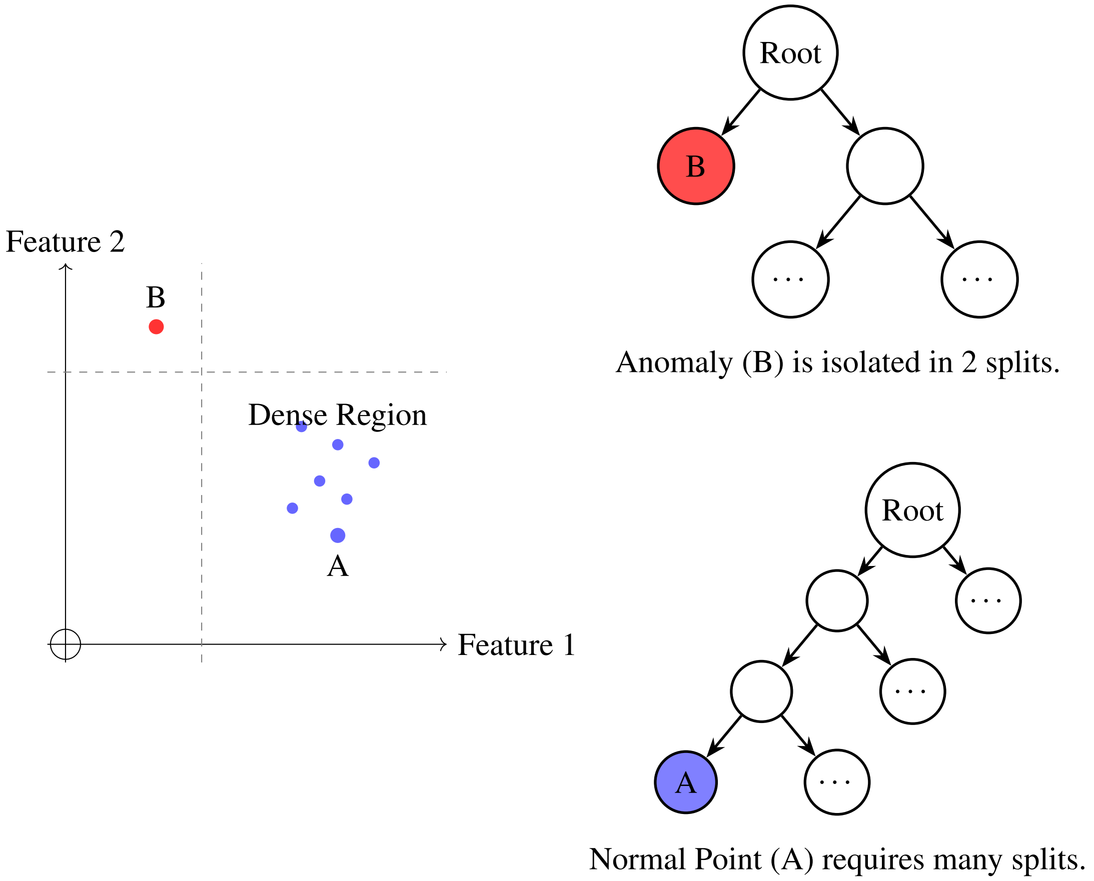
Chapter 5

The Unsupervised Risk Model: Isolation Forest

5.1 Algorithmic Deep Dive

5.1.1 Core Intuition

The fundamental principle of Isolation Forest is that **anomalies are easier to isolate than normal points**. The model builds a forest of random decision trees. To ”isolate” a data point, the algorithm randomly selects a feature and then a random split point. This process is repeated until the point is alone in a node. Anomalous points require fewer splits to be isolated.



Anomaly (B) is isolated in 2 splits.

Normal Point (A) requires many splits.

Figure 5.1: A visual intuition for Isolation Forest. A point (B) in a sparse region is isolated with few splits (short path), while a point (A) in a dense region requires many splits (long path).

5.1.2 The Mathematics of Anomaly Scoring

The algorithm quantifies this ”ease of isolation” into a formal anomaly score.

Path Length $h(x)$: For a given data point x, the path length $h(x)$ is the number of edges it traverses from the root of an iTree to a terminating leaf node.

Expected Path Length $E(h(x))$: Since the process is random, the algorithm builds a forest of iTrees and calculates the expected (average) path length $E(h(x))$ for point x across all trees.

Normalization Constant $c(n)$: $E(h(x))$ is normalized by $c(n)$, the average path length for a search in a Binary Search Tree given n data points. It is defined as:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}$$

where n is the number of instances and $H(i)$ is the harmonic number.

Final Anomaly Score $s(x,n)$: The final score is an exponential function of the normalized expected path length:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

This formula produces a score s between 0 and 1:

- If $E(h(x)) \to 0$ (very short path), then $s \to 1$. This is a clear **anomaly**.
- If $E(h(x)) \to n-1$ (very long path), then $s \to 0$. This is a clear **normal** point.
- If $E(h(x)) \approx c(n)$, then $s \approx 0.5$. The point is not clearly normal or anomalous.

In practice, we use the model's `decision_function` score, which is related to $-E(h(x))$.

Chapter 6

Model Explainability: A Deep Dive into SHAP

6.1 The Imperative for Transparency in FinTech

In financial applications, a ”black box” model is unacceptable. For regulatory reasons (e.g., GDPR’s ”right to explanation”), for fairness auditing, and for business trust, it is essential to explain why a model made a specific decision.

6.2 Deconstructing the Waterfall Plot

The primary visualization used in the dashboard is the SHAP waterfall plot. It provides a complete breakdown of a single prediction:

- $E[f(X)]$ **(The Base Value):** The average model output over the training data.
- **Feature Contributions (The Bars):** Each bar shows how a feature’s value pushed the prediction away from the base value. Red bars push the score **up** (more normal, less risky), while blue bars push the score **down** (more anomalous, riskier).
- $f(X)$ **(The Final Value):** The final raw anomaly score for the user, which is the sum of the base value and all contributions.

Chapter 7

Results, Analysis Dashboard

7.1 Final Scoring Pipeline

The journey from raw data to a final score is a multi-step pipeline designed to be both mathematically robust and logically sound.

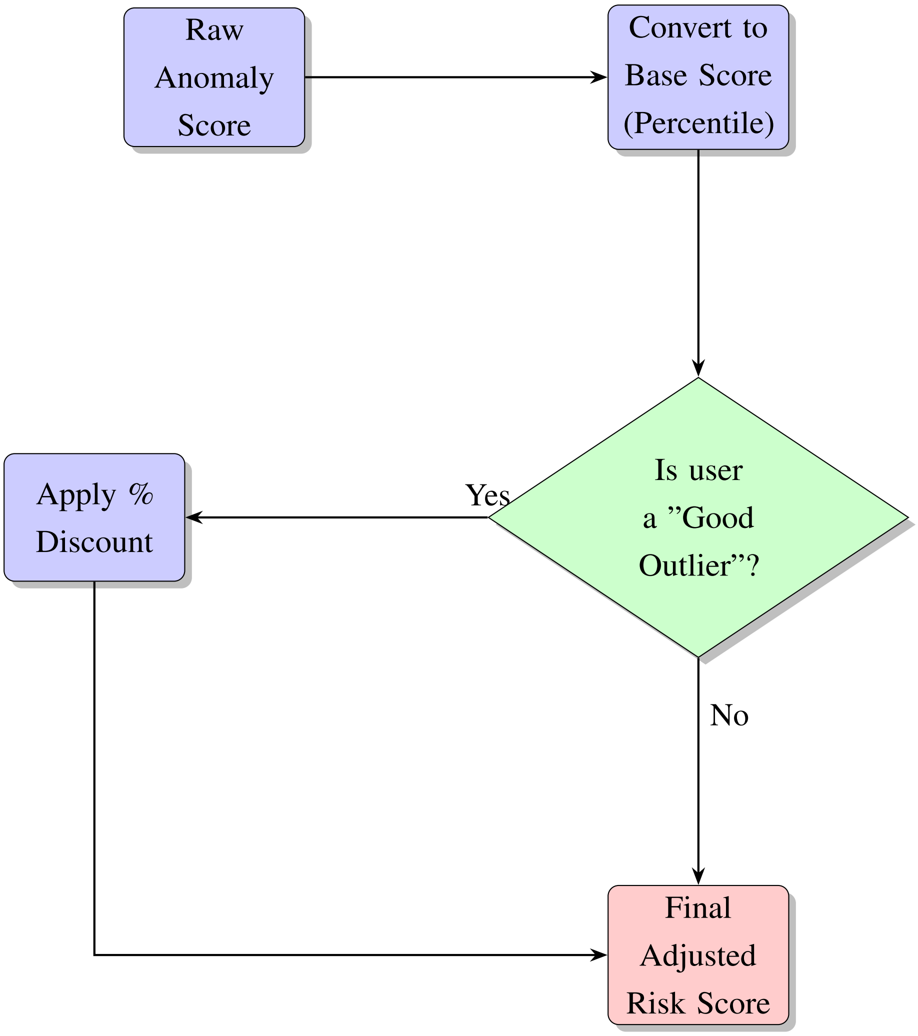

Raw Anomaly Score
Convert to Base Score (Percentile)
Is user a ”Good Outlier”?
Yes
Apply % Discount
No
Final Adjusted Risk Score

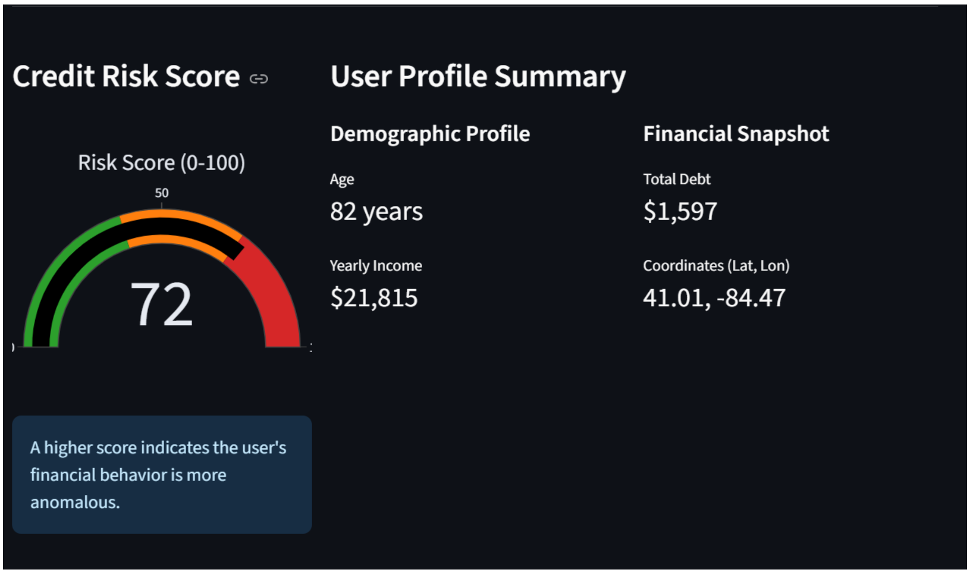
Figure 7.1: The Final Scoring Pipeline, with an explicit logical adjustment step for ”good outliers.”

7.2 Analysis of Results

The following case studies demonstrate the model’s ability to handle diverse user profiles, from clear high-risk cases to nuanced ”good outliers.”

7.2.1 Case Study 1: A High-Risk User Profile (User 80)

This user represents a profile with clear and significant risk factors that the model correctly identifies.



Credit Risk Score

User Profile Summary

Risk Score (0-100)

A higher score indicates the user's financial behavior is more anomalous.

Demographic Profile

Age
82 years

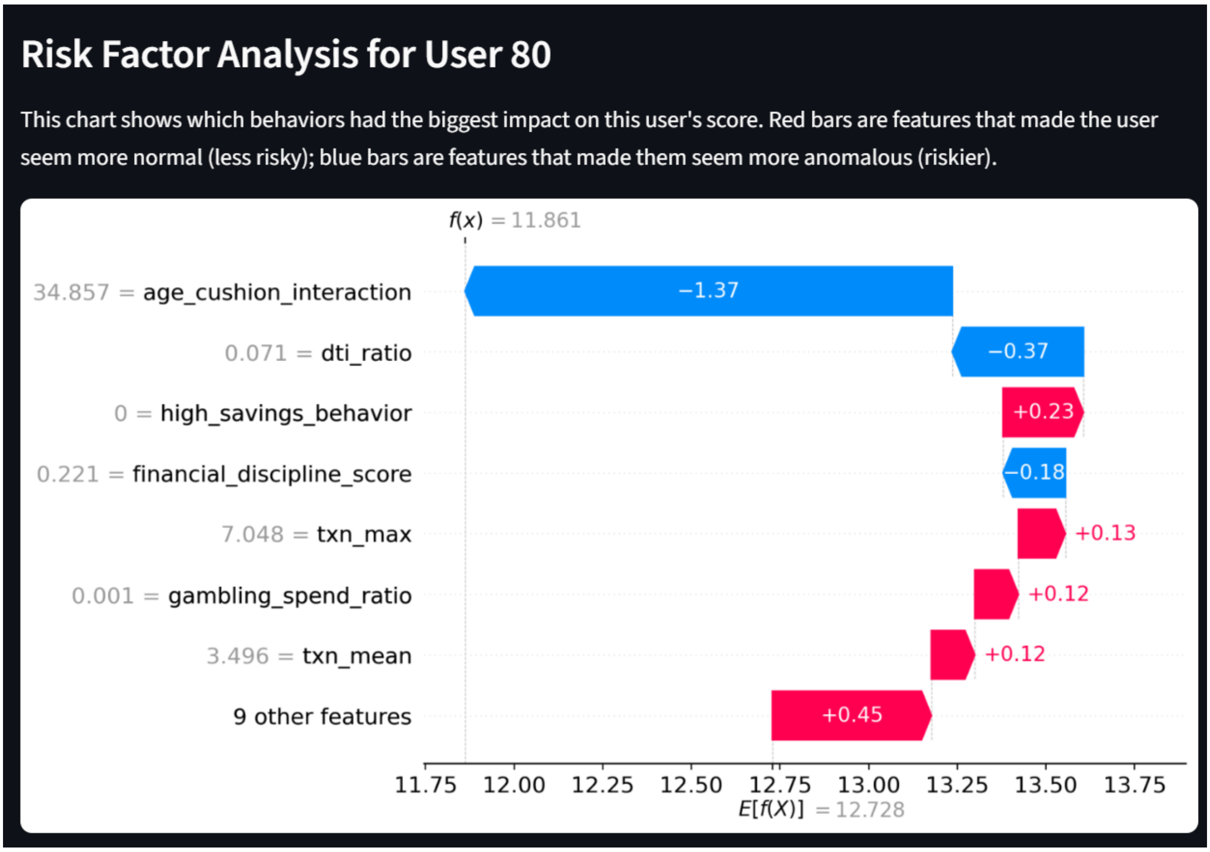
Yearly Income
$21,815

Financial Snapshot

Total Debt
$1,597

Coordinates (Lat, Lon)
41.01, -84.47

Figure 7.2: Dashboard view for User 80, who was assigned a high-risk score of 72.



Risk Factor Analysis for User 80

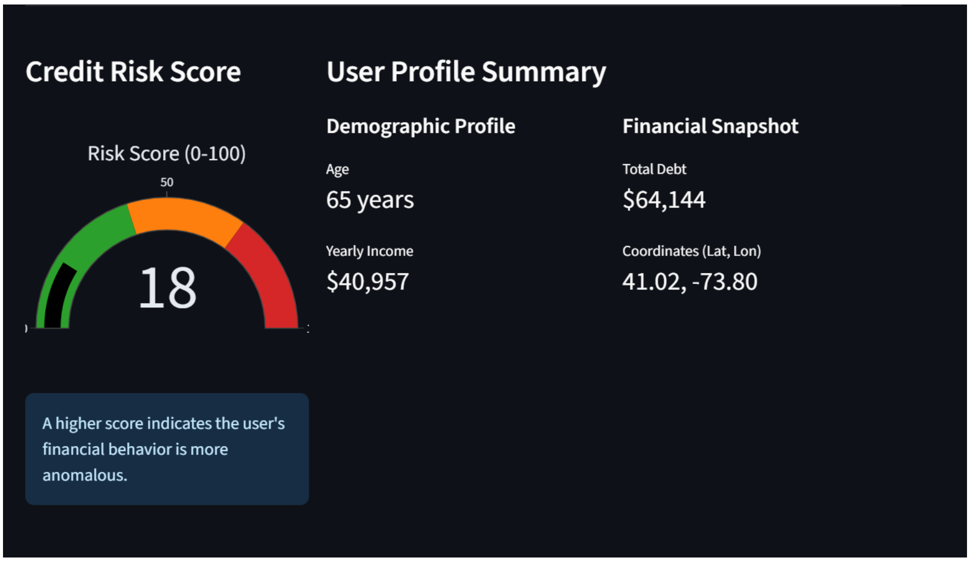
This chart shows which behaviors had the biggest impact on this user's score. Red bars are features that made the user seem more normal (less risky); blue bars are features that made them seem more anomalous (riskier).

Figure 7.3: SHAP Waterfall Plot for User 80, detailing the drivers of the high-risk score.

Analysis: User 80 was assigned a final risk score of 72. This high-risk assessment is primarily driven by demographic and financial leverage factors, as shown in their SHAP plot. The single largest negative contributor is the `age_cushion_interaction`, indicating that the user's advanced age (82) combined with a weak financial cushion creates a significant risk signal. The `dti_ratio` also contributes negatively, pushing the score further into anomalous territory. The model correctly identifies these compounding factors and assigns a high risk score.

7.2.2 Case Study 2: A Low-Risk ”Good Outlier” (User 460)

This user is an excellent example of the logical adjustment engine at work. While their profile is statistically anomalous, their underlying financial health is strong.



Credit Risk Score

A higher score indicates the user's financial behavior is more anomalous.

User Profile Summary

Demographic Profile

Age
65 years

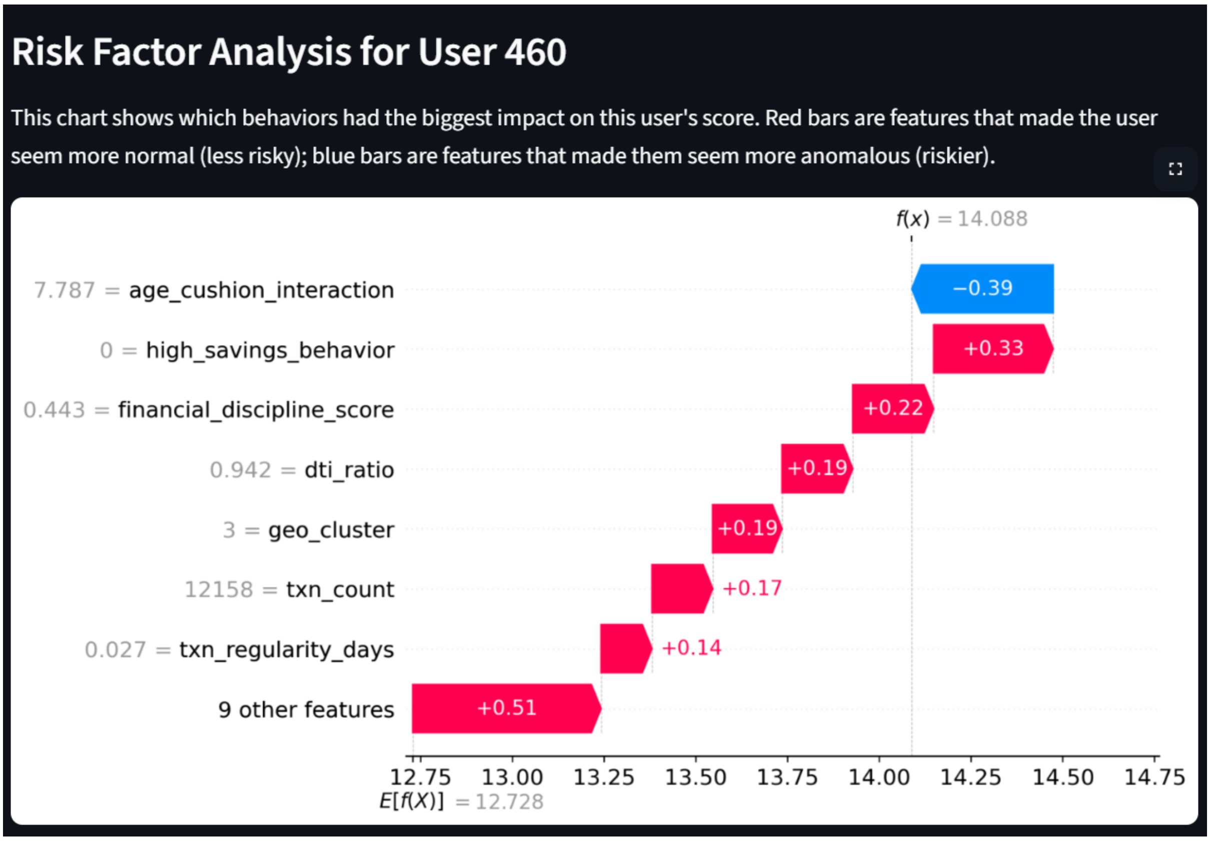
Yearly Income
$40,957

Financial Snapshot

Total Debt
$64,144

Coordinates (Lat, Lon)
41.02, -73.80

Figure 7.4: Dashboard view for User 460, who received a final low-risk score of 18.



Risk Factor Analysis for User 460

This chart shows which behaviors had the biggest impact on this user's score. Red bars are features that made the user seem more normal (less risky); blue bars are features that made them seem more anomalous (riskier).

Figure 7.5: SHAP Waterfall Plot for User 460. Note the mix of positive (red) and negative (blue) factors.

Analysis: User 460 received a final score of 18, indicating very low risk. The SHAP plot reveals a more complex story. The model flags their `age_cushion_interaction` as a risk factor (blue bar) due to their age (65) and high total debt. However, this is counteracted by strong positive signals (red bars), including a high `financial_discipline_score`. Most importantly, although their raw anomaly score was high, the logical adjustment engine identified strong positive attributes (like high savings behavior) and applied a significant discount, resulting in a final score that accurately reflects their responsible financial management.

7.3 Global Feature Importance

Beyond individual cases, the global SHAP summary shows which features have the most predictive power across the entire user base.

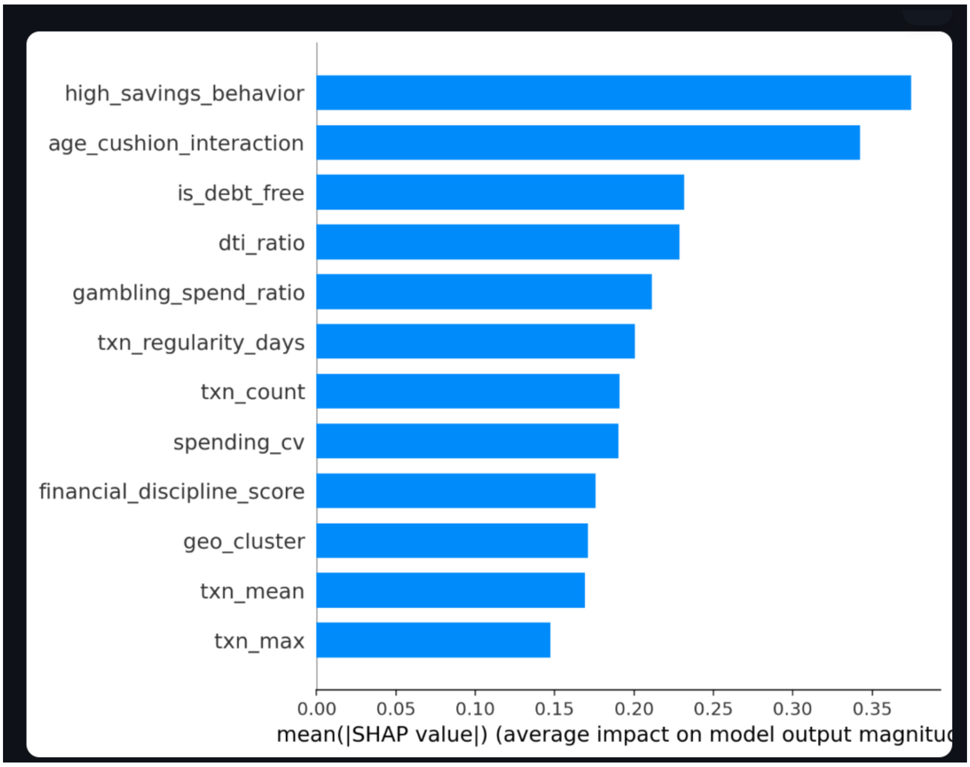


Figure 7.6: Global Feature Importance, showing the average impact of each feature on the model's output.

Analysis: The global importance chart reveals the core logic of the model. The two most impactful features are `high_savings_behavior` and the `age_cushion_interaction`. This confirms that the model's decisions are driven by a sophisticated understanding of both positive ”golden features” (like a user's ability to save) and compounded risk factors (the interplay between age and financial stability). Other key drivers include debt-related metrics and behavioral signals, validating the advanced feature engineering approach.

Chapter 8

Conclusion

8.1 Summary of Achievements

The FinShield project successfully produced a high-quality, unsupervised credit risk model. By focusing on advanced behavioral features and implementing a logical adjustment layer, the system provides objective, transparent, and scalable risk scores ideally suited for financial inclusion initiatives.

8.2 Future Work

While robust, the model can be enhanced via more sophisticated behavioral and architectural patterns.

- **Deeper Time-Series Analysis:** Implement features like the **slope of monthly spending** (via linear regression) to capture a user's financial trajectory, distinguishing stable users from those with increasing lifestyle inflation.
- **Unsupervised Behavioral Clustering:** Before feature engineering, use K-Means on key metrics to automatically discover financial ”personas.” The resulting cluster ID (”Frugal Planner,” ”Leveraged Spender”) would serve as a single, powerful categorical feature for the main model.
- **Two-Stage Hybrid Model Architecture:** For ultimate performance, a two-stage model could be implemented.
  - **Stage 1:** The Isolation Forest runs as is, producing a raw `anomaly_score`.
  - **Stage 2:** A supervised model (e.g., XGBoost) is trained to predict an expert-defined **proxy label** (a ”Desirable Profile Score” created from the Golden Features). This model would use the original features *plus* the `anomaly_score` from Stage 1 as its input, combining unsupervised discovery with expert-guided learning.