VPBank Technology Hackathon 2025

General Brief

Please fill up this table and use this document as a template to write your proposal.

Challenge Statement	**AI/ML Enhance Credit Scoring**
Team Name	Team 255

Team Members

Full Name	**Role**	**Email Address**	**School Name (if applicable)**	**Faculty / Area of Study**	**LinkedIn Profile URL**
Nguyễn Đăng Huy	Leader	Huy2492002@gmail.com	N/A	Computer Science - Data Science & AI	N/A
Nguyễn Nhật Minh	Member	minh.nn0402@gmail.com	N/A	Computer Science - Data Science & AI	https://www.linkedin.com/in/nhatminh432/
Đỗ Thành Đức	Member	do.thanh.duc.2002@gmail.com	SOICT - HUST	Computer Science - Data Science & AI	N/A
Trần Trọng Khang	Member	Ttkhang202@gmail.com	N/A	Computer Science - Data Science & AI	N/A

Content Outline

Solutions Introduction

What is Our Solution and How Does it Work?

Our "AI/ML Enhanced Credit Scoring" solution is a cutting-edge system designed to **revolutionize credit assessment** by leveraging Artificial Intelligence (AI), Machine Learning (ML), and Large Language Models (LLMs). Unlike traditional systems that rely solely on structured financial data, our solution incorporates diverse **unconventional data sources** (e.g., transaction behavior, utility bills, e-commerce activity) to provide a holistic view of creditworthiness, especially for new-to-bank or thin-file customers.

The system offers a **continuous, adaptive workflow** from data ingestion to model deployment and explainable scoring.

Main Features:

- **Multi-source Data Ingestion:** Handles both structured and unstructured data inputs.
- **Intelligent Feature Engineering:** Utilizes AWS Glue and Amazon Textract to process raw data into valuable features.
- **Centralized Feature Store:** SageMaker Feature Store ensures consistent and reusable features for training and inference.
- **Automated Model Lifecycle:** SageMaker Pipelines orchestrates training, evaluation, and deployment of ML models (e.g., XGBoost, LightGBM) and integrates LLM capabilities.
- **Real-time & Batch Scoring:** Provides low-latency real-time scoring via API Gateway and Lambda, alongside scalable batch processing.
- **LLM-powered Explainability:** Generates clear, natural language explanations for credit scores, enhancing transparency.
- **Proactive Monitoring & Adaptation:** CloudWatch detects model/data drift, triggering re-training via SageMaker Pipelines to maintain accuracy.
- **User-Friendly Interface:** An intuitive frontend for seamless user interaction and report viewing.
- **Automated CI/CD:** Ensures rapid and reliable model and application deployments via GitHub and AWS Code Pipeline.

Impact of Solution

How Does Your Solution Benefit Society / The Target Audience?

Our solution significantly benefits financial institutions and customers by:

- **Promoting Financial Inclusion:** By analyzing non-traditional data, it enables access to credit for previously underserved "new-to-bank" or "thin-file" customers.
- **Reducing Risk & Increasing Efficiency:** Improved accuracy in risk assessment leads to better lending decisions, reduced bad debt, and optimized credit portfolios for financial institutions.
- **Enhancing Transparency & Fairness:** Natural language explanations empower customers to understand credit decisions, fostering trust and fairness, and aiding regulatory compliance.
- **Adapting to Diverse Customer Segments:** Designed to flexibly score various customer types (e.g., salaried, freelancers), reflecting their unique financial profiles.

Why Is Your Solution a Good Solution? How Is It Better Than Existing Solutions / Competitors?

Our solution stands out due to:

- **Comprehensive Non-Traditional Data Integration:** Unlike many traditional systems, we fully leverage diverse unstructured data sources for a 360-degree customer view.
- **Superior Explainability via LLM:** A key differentiator, our solution doesn't just provide a score but explains *why* a score was assigned in clear, natural language, a challenge for conventional ML models.
- **Seamless, End-to-End AWS Integration:** Built entirely on AWS managed services, the entire ML lifecycle—from ingestion to monitoring—is automated and tightly integrated, optimizing performance and reducing manual overhead.
- **Adaptive & Continuously Improving:** Automated training pipelines and robust monitoring ensure the system learns and adapts to new data trends, maintaining accuracy over time.

What Is Your Solution’s Competitive Advantage / Unique Selling Point?

Our unique selling proposition (USP) is the **powerful combination of traditional AI/ML with Generative AI (LLMs) to deliver highly accurate, comprehensive, and inherently explainable credit scoring.**

While competitors may focus on one aspect (ML for scoring or LLM for explanation), our solution synergistically integrates both to:

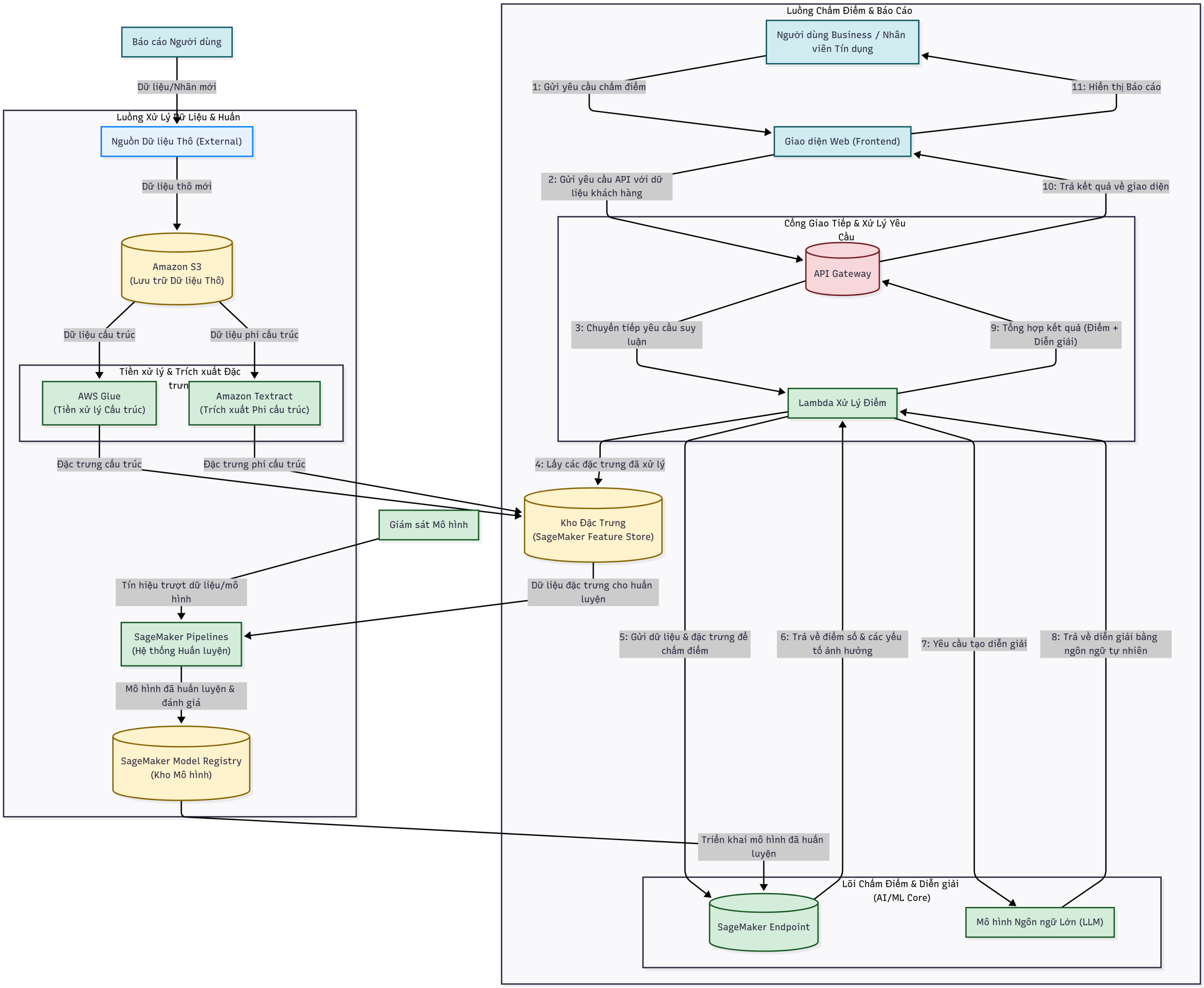
1. **Achieve superior credit risk assessment** by utilizing diverse, often overlooked data.
2. **Provide unparalleled transparency** through natural language explanations, building trust and acceptance.

This creates a system that is not only more effective at-risk assessment but also more user-friendly, trustworthy, and fair.

Deep Dive into Solution

Detailed Write-up of Solution (Level 2 Data Flow Diagram)

Our solution follows a rigorous data processing and model workflow, ensuring a seamless journey from raw data to credit score and explanation. The detailed Level 2 Data Flow Diagram (DFD) below illustrates the interactions between system components:


Báo cáo Người dùng
Dữ liệu/Nhãn mới
Luồng Xử Lý Dữ Liệu & Huấn
Nguồn Dữ liệu Thô (External)
Dữ liệu thô mới
Amazon S3
(Lưu trữ Dữ liệu Thô)
Dữ liệu cấu trúc
Dữ liệu phi cấu trúc
Tiền xử lý & Trích xuất Đặc trưn
AWS Glue
(Tiền xử lý Cấu trúc)
Amazon Textract
(Trích xuất Phi cấu trúc)
Đặc trưng cấu trúc
Đặc trưng phi cấu trúc
Giám sát Mô hình
Tín hiệu trượt dữ liệu/mô hình
SageMaker Pipelines
(Hệ thống Huấn luyện)
Mô hình đã huấn luyện & đánh giá
SageMaker Model Registry
(Kho Mô hình)
Luồng Chấm Điểm & Báo Cáo
Người dùng Business / Nhân viên Tín dụng
1: Gửi yêu cầu chấm điểm
11: Hiển thị Báo cáo
Giao diện Web (Frontend)
2: Gửi yêu cầu API với dữ liệu khách hàng
10: Trả kết quả về giao diện
Cổng Giao Tiếp & Xử Lý Yêu Cầu
API Gateway
3: Chuyển tiếp yêu cầu suy luận
9: Tổng hợp kết quả (Điểm + Diễn giải)
Lambda Xử Lý Điểm
4: Lấy các đặc trưng đã xử lý
Kho Đặc Trưng
(SageMaker Feature Store)
Dữ liệu đặc trưng cho huấn luyện
5: Gửi dữ liệu & đặc trưng để chấm điểm
6: Trả về điểm số & các yếu tố ảnh hưởng
7: Yêu cầu tạo diễn giải
8: Trả về diễn giải bằng ngôn ngữ tự nhiên
Triển khai mô hình đã huấn luyện
Lõi Chấm Điểm & Diễn giải
(AI/ML Core)
SageMaker Endpoint
Mô hình Ngôn ngữ Lớn (LLM)

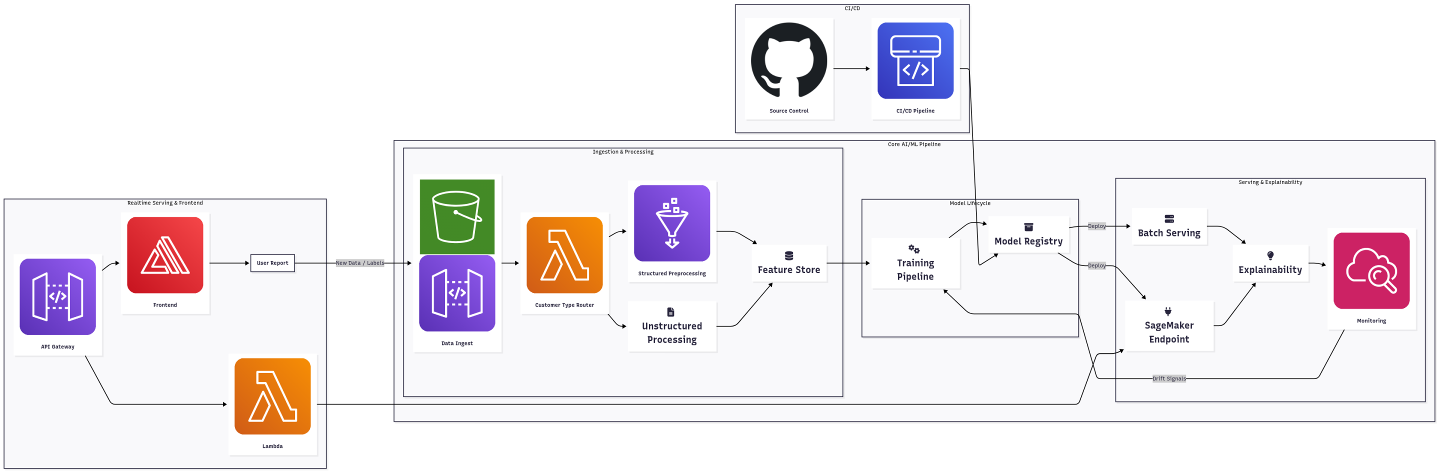
Key Flows:

- **Offline/Batch Flow:** Raw data from external sources is stored in **Amazon S3**. **AWS Glue** processes structured data, while **Amazon Textract** extracts information from unstructured data. Both feed processed features into **SageMaker Feature Store**. **SageMaker Pipelines** orchestrates model training using these features, registering the best models in **SageMaker Model Registry**. **Model Monitoring** (CloudWatch) detects drift and triggers re-training via Pipelines. User feedback provides new data for continuous improvement.
- **Real-time Scoring Flow:** Business users interact via a **Web Interface**, sending requests through **API Gateway**. A **Lambda Function** retrieves features from Feature Store and sends data to a **SageMaker Endpoint** for scoring. The Lambda then uses an **LLM** to generate natural language explanations for the score. The combined score and explanation are returned to the user via API Gateway and the Web Interface.

Architecture of Solution

How Does Your Solution Make Use of AWS Infrastructure?

Our solution is built entirely on **Amazon Web Services (AWS)**, leveraging its managed services for scalability, reliability, security, and cost-effectiveness. We integrate these services seamlessly to create a robust and high-performing AI/ML credit scoring system.



AWS Service Utilization & Integration:

- **Data Ingestion & Storage: Amazon S3** acts as our scalable data lake for all raw data. **API Gateway** provides secure API endpoints for data submission, triggering **AWS Lambda** to process and store data in S3.

- **Data Processing & Feature Engineering: AWS Lambda** intelligently routes incoming data based on customer type. **AWS Glue** performs ETL on structured data, while **Amazon Textract** extracts information from unstructured documents. All processed features are stored in **Amazon SageMaker Feature Store**, ensuring consistency for both training and real-time inference.
- **Model Training & Management: Amazon SageMaker Pipelines** orchestrates the entire ML workflow, from feature retrieval from Feature Store to model training and evaluation. Trained models are registered in **Amazon SageMaker Model Registry** for version control and deployment management.
- **Model Serving:**
    - **Real-time: API Gateway** receives scoring requests. A dedicated **AWS Lambda** function fetches features from Feature Store and invokes **Amazon SageMaker Endpoint** (where the core ML model is deployed) for inference. The Lambda then uses an **LLM** (potentially another SageMaker Endpoint or external API) for natural language explanations, aggregating results before returning them via API Gateway.
    - **Batch:** For large-scale offline scoring, **SageMaker Batch Transform** processes data directly from S3, applies the model, and stores results back in S3.
- **Explainability:** An **LLM** (Large Language Model) is integrated (via a dedicated SageMaker Endpoint or external API, orchestrated by Lambda) to provide human-readable explanations for the generated credit scores.
- **Monitoring: Amazon CloudWatch** collects metrics and logs from all services. It monitors model performance (e.g., data/model drift) and triggers **SageMaker Pipelines** to re-train models if drift is detected, ensuring continuous accuracy.
- **Frontend: AWS Amplify** is used to build and host the interactive web interface, which seamlessly communicates with the API Gateway for user interactions.
- **CI/CD: GitHub** serves as the source control repository. **AWS CodePipeline** automates the entire CI/CD process, triggering builds (via CodeBuild) and deployments (to Lambda, SageMaker, etc.) upon code changes, ensuring rapid and reliable delivery of new features and model versions.

This architecture leverages AWS's fully managed services to deliver a scalable, secure, and highly efficient AI/ML credit scoring solution with built-in explainability.