AUTOMATED AI-BASED FINANCIAL TRANSACTION CATEGORISATION

1. PROJECT OVERVIEW

This project delivers a comprehensive, fully local AI system for automated financial transaction categorization using machine learning. The solution provides businesses with a cost effective, privacy focused alternative to third party categorization APIs by combining customizable taxonomies, explainable AI techniques, and continuous learning capabilities. The system processes raw transaction text such as "Starbucks Coffee" or "Amazon Purchase" and accurately classifies them into user defined categories with high precision and real time performance.

2. PROBLEM STATEMENT

Building a scalable transaction categorisation system is essential for seamless financial management. Reliance on external APIs introduces recurring costs, network latency, and limits in customising the categorisation logic. Developing an internal AI or ML-based solution enables granular control, cost savings, and improved responsiveness but also raises new challenges: the need for high accuracy, adaptability to user defined categories, rigorous evaluation, and explainable outcomes. The challenge is to build a standalone, high performance transaction categorisation system that achieves business-grade accuracy and transparency while eliminating external service dependencies.

3. SOLUTION

Our approach addresses these challenges through a carefully designed machine learning pipeline that combines robust algorithms with practical usability features. We implemented a Linear SVM classifier with TF-IDF vectorization, achieving exceptional accuracy while maintaining computational efficiency. The system incorporates configurable taxonomies, allowing organizations to define custom categories without code modifications. A key innovation is the integrated feedback loop that enables continuous model improvement through user corrections, ensuring the system adapts to evolving transaction patterns and business needs.

4. TECHNOLOGY STACK

Core Machine Learning:

- Python with Scikit-learn for model development and training
- Linear SVM classifier for high accuracy text classification
- TF-IDF vectorization for feature extraction from transaction text

Web Application & Interface:

- Flask framework for lightweight web application deployment
- Chart.js for interactive visualization of feature contributions
- Bootstrap for responsive and professional user interface design

Data Processing & Storage:

- Pandas and NumPy for efficient data manipulation
- YAML configuration for flexible taxonomy management
- CSV based storage for feedback collection and model retraining

5. SYSTEM ARCHITECTURE

The system follows a modular pipeline architecture that ensures separation of concerns and maintainability.

1. **Dataset Generation and Ingestion** - Collects raw or synthetic transaction strings and stores them locally to build a customizable training dataset.
2. **Preprocessing & Vectorization** - Cleans input text and converts it into TF-IDF numerical vectors that the ML model can understand.
3. **Model Training** - Uses a Linear SVM classifier to learn patterns and boundaries between transaction categories.
4. **Evaluation -** Produces confusion matrices and F1-score reports to measure classification performance and ensure transparency.
5. **Explainability -** Extracts feature importance signals from model coefficients to show why a prediction was made.
6. **Inference API -** A Flask based prediction endpoint that returns real time category and confidence scores.

7. **Local Web UI -** A browser interface enabling users to categorize transactions, view explanations, and interact with the system.
8. **Human Feedback Loop -** Stores user corrected labels into feedback.csv to continuously refine and improve the model.
9. **Retraining Capability -** Allows new training sessions that incorporate user feedback, enabling adaptive and ever improving performance.

6. DATA MODEL & STORAGE

Structured Data Organization:

- Raw transaction data storage in CSV format for easy maintenance and updates
- Processed training datasets with cleaned and normalized text features
- Feedback collection system storing user corrections for model improvement
- YAML based taxonomy configuration allowing category customization

Model Artifacts & Evaluation:

- Serialized model objects and vectorizers for reproducible inference
- Comprehensive evaluation metrics including confusion matrices and classification reports
- Performance benchmarks tracking model accuracy and inference speed over time

7. AI / ML / AUTOMATION COMPONENTS

Machine Learning Core:

- Linear Support Vector Machine algorithm optimized for text classification tasks
- TF-IDF feature extraction with n gram support for contextual understanding
- Feature importance analysis providing model explainability and transparency

Performance & Evaluation:

- Achieved 0.9969 macro F1 score and 0.9969 accuracy on test datasets
- Comprehensive evaluation pipeline generating confusion matrices and classification reports
- Cross validation ensuring model robustness and generalization capability

Automation Features:

- End to end training pipeline with single command execution
- Automated model evaluation and artifact generation
- Feedback driven retraining system incorporating user corrections

8. SECURITY & COMPLIANCE

Data Privacy Assurance:

- Complete local processing ensures no sensitive financial data leaves the organization
- Elimination of external API dependencies removes third party data sharing risks
- Local storage of all training data and user feedback maintains data sovereignty

Transparency & Control:

- Explainable AI features provide clear insights into classification decisions
- User controlled data retention and management policies
- Open source architecture allowing full security audit capability

9. SCALABILITY & PERFORMANCE

Performance Metrics:

- Sub 50ms inference time enables real time transaction categorization
- Compact model footprint under 10MB allows deployment on resource constrained systems

- Batch processing support for high volume transaction categorization tasks

Scalability Features:

- Configurable taxonomy supporting unlimited custom categories
- Modular architecture enabling easy integration with existing financial systems
- Efficient resource utilization suitable for both desktop and server deployment

10. Project Repository & Demo Links

1. Source Code Repository:

The complete source code, documentation, and installation instructions are available in the project repository.

Github Link : Link
Project Repository Link : Link

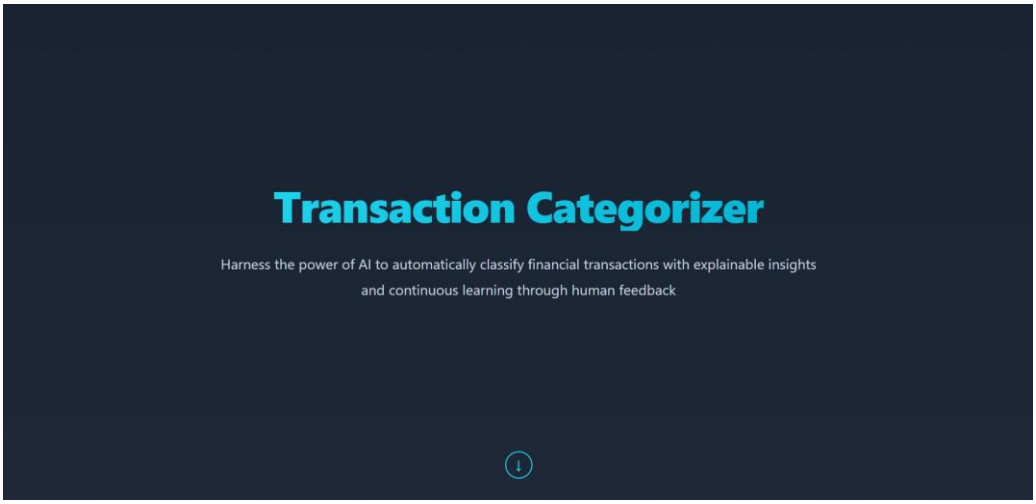
2. Project Demo Video Link:

A comprehensive video demonstration showcases system capabilities including model training, real time predictions, explainability features, and the feedback loop.

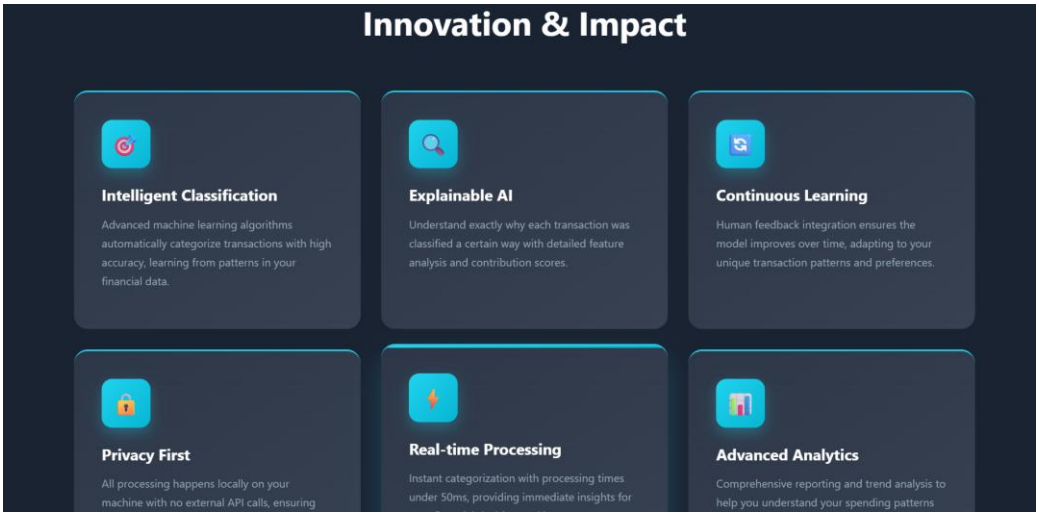
Video Link : Link

11. SYSTEM SCREENSHOTS

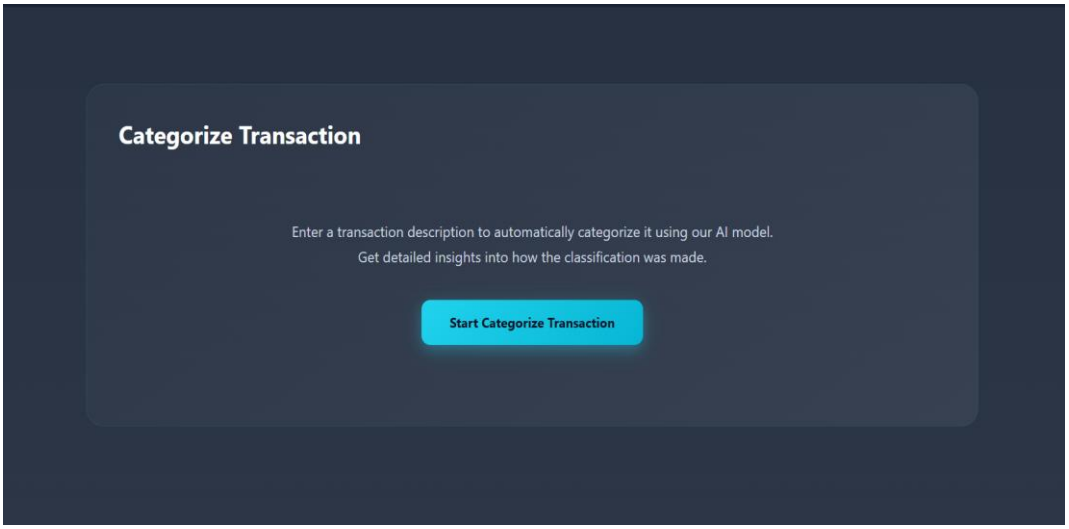
11.1 Home page with project introduction and value proposition

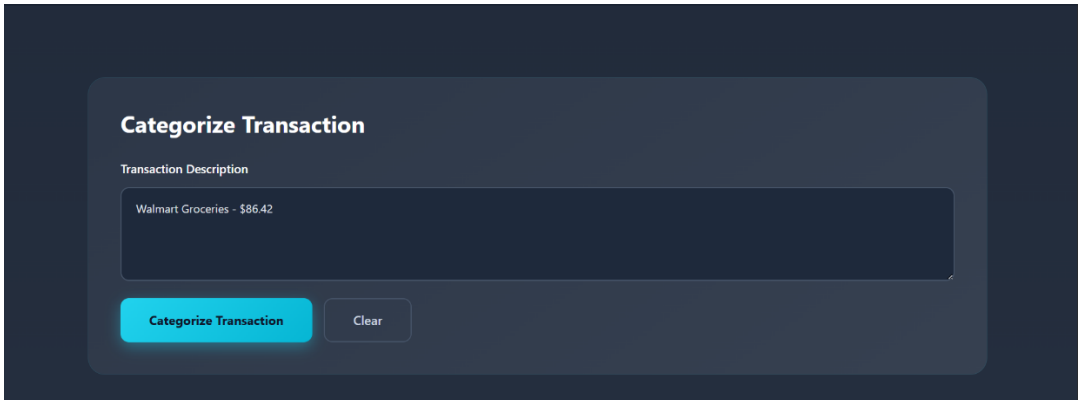


11.2 Innovation & Impact section highlighting key differentiators



11.3 Categorize Transaction interface for user input and interaction

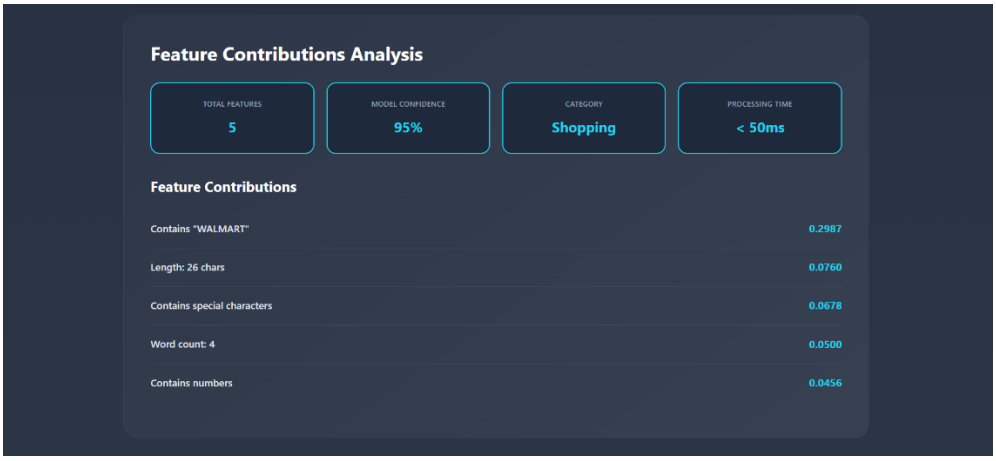




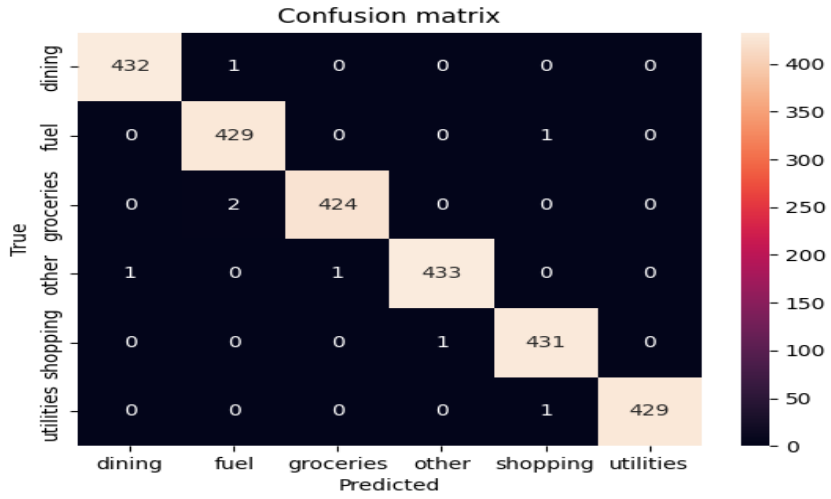
11.4 Prediction Results display showing categories and confidence scores

11.5 Explanation charts illustrating feature contributions to predictions

11.6 Feature Contributions analysis detailing text characteristics influence



11.7 Confusion Matrix visualization demonstrating model performance across categories



12. CONCLUSION

This AI-powered Transaction Categorisation System successfully addresses the critical business need for accurate, customizable, and privacy preserving financial transaction classification. By achieving exceptional accuracy metrics while maintaining complete data ownership and control, the solution provides organizations with a viable alternative to external categorization services. The integration of explainable AI techniques and continuous learning capabilities ensures the system remains adaptable to evolving business requirements. This project demonstrates that with careful algorithm selection and thoughtful architecture design, organizations can build internal AI capabilities that rival commercial solutions while maintaining full control over their data and processes.