MIT-WPU
॥ विश्वशान्तिर्ध्रुवं ध्रुवा ॥

Since 1983
Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

MIT tbi

सत्यमेव जयते
Ministry of Education
Government of India

All India Council for Technical Education
AICTE

INSTITUTION'S INNOVATION COUNCIL
(Ministry of Education Initiative)

HACK

MIT-WPU'2

MIT-WPU' 25

ले छलांग

ONE HACKATHON. THIRTEEN TRACKS.

MILLION PROBLEMS. BILLION MINDS.

NEURALS

Sr.No	NAME	PHONE	EMAIL ID:	ROLE
1	-	-	-	Faculty Mentor
2	Mihir Mohite	80079 53471	mihir.mohite@mitwpu.edu.in	Team Leader
3	Shreyash Dhoot	76665 97668	shreyash.dhoot@mitwpu.edu.in	Team Member 1
4	Nishad Dere	91724 99079	nishad.dere@mitwpu.edu.in	Team Member 2
5	-	-	-	Team Member 3

PROBLEM STATEMENT:

Bank A aims to enhance its risk management framework for existing Credit Card customers by developing a **Behavior Score**—a predictive model that estimates the probability of future defaults. The model will be trained on active, non-delinquent customers and leverage advanced ML techniques to assess risk. The Behavior Score will be utilized for various portfolio risk management activities, ensuring proactive risk mitigation and sustained profitability.

Problem Understanding and Objective

Real-World Context:

- Rising Defaults: U.S. credit card defaults reached record levels in 2024, with approximately $46B in write-offs—the highest seen since 2010. (Financial Times)
- Economic Ripple Effects: High default rates lead to increased bad debt, reduced bank liquidity, and can trigger broader financial stress, as evidenced in past crises. (NYPOST, Wiki.org)

Why It's Important:

- Financial Stability: Elevated default rates stress banks' capital and hinder their ability to lend, risking systemic issues and higher borrowing costs for businesses.
- Profitability at Stake: Ineffective risk management can result in significant financial losses and reduced investor confidence, potentially impacting the entire credit market.
- Data-Driven Decision Making: Advanced predictive models (using machine learning, credit scoring) have proven to reduce default rates by identifying high-risk customers early.

Our Objective – What We're Predicting/Solving:

- Predict Future Defaults: Develop a "Behavior Score" that estimates the probability of future defaults among active, non-delinquent credit card customers.
- Proactive Risk Mitigation: Utilize the score for dynamic portfolio management, enabling timely interventions to minimize losses and sustain profitability.

DATA CLEANING AND PREPROCCESSING

Missing Value Treatment:
- Dropped columns with >50% missing values
- Imputed remaining numeric variables using the median (minimizes outlier influence)

Handling Invalid Data / Anomalies:
- Median imputation mitigated the impact of anomalies and prevented skewed distributions

Encoding, Scaling & Transformation:
- No encoding required (all features numeric)
- StandardScaler normalized feature values

Feature Transformation:
- Transformed onus_attribute_1 into a binary indicator (e.g., onus_attribute_below_500k) since most defaulters fall below this threshold

Train/Test Consistency:
- Stratified split maintained target imbalance
- Applied identical imputation and feature-dropping across both sets
- Reindexed test set to match training features

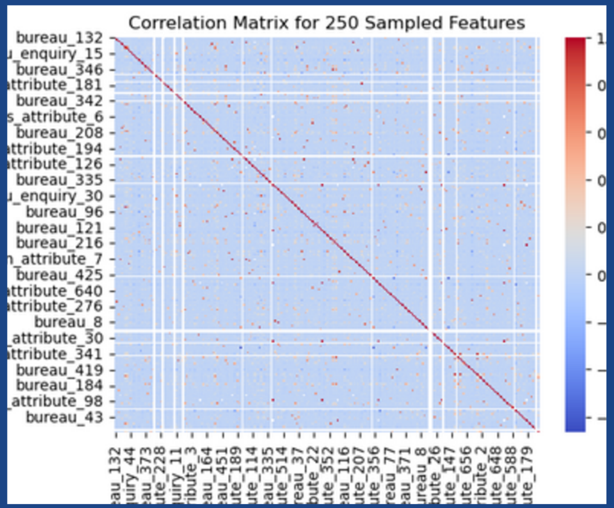
Feature Reduction & Selection:
- Dropped redundant features (|corr| > 0.95) by retaining the one with stronger correlation to the target
- Removed features with near-zero target correlation (|corr| < 0.017) to reduce noise

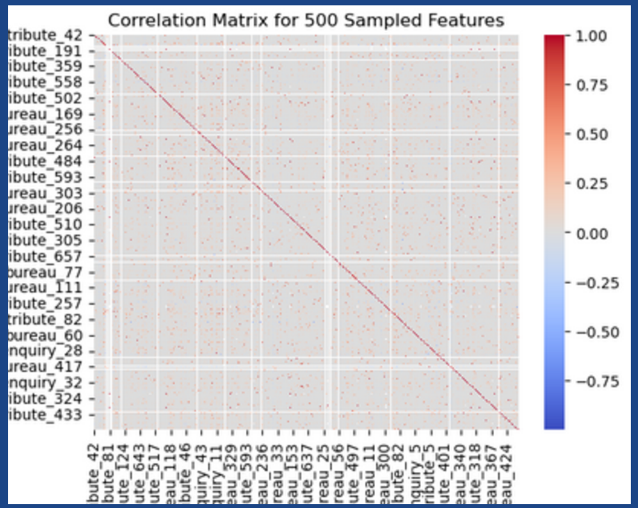
Non-linear Insights:
- Leveraged mutual information (via GPU acceleration) alongside Pearson correlation
- Flagged features with high MI – |Pearson| differences, revealing hidden non-linear relationships

EXPLORATORY DATA ANALYSIS(EDA)

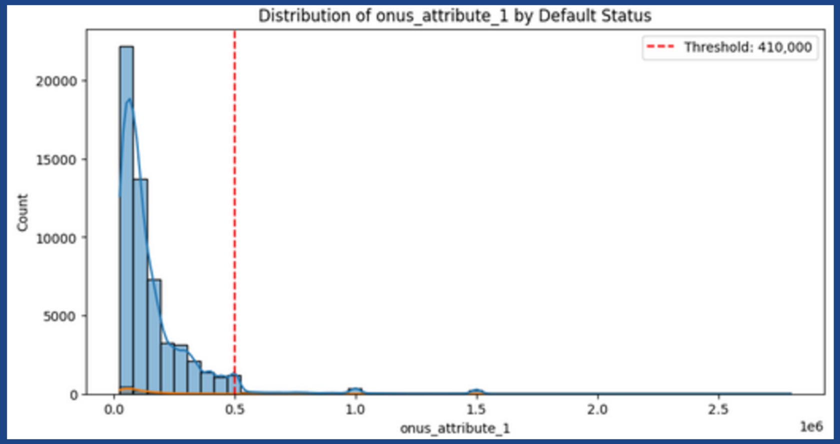
Randomised Feature Correlation: 250 v/s 500 (total 1216)

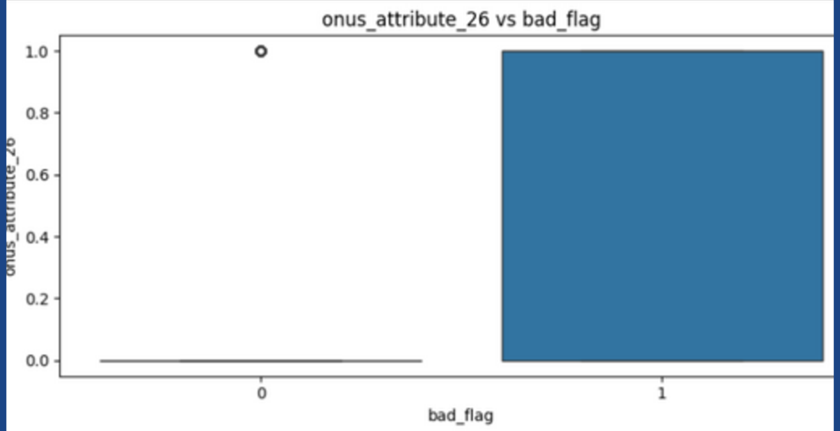
- To judge the cross-correlation between data, randomly plotted 250 and 500 features, the trend suggested that the entire feature set was mostly uncorrelated.
- Also many features were highly redundant (with absolute correlations exceeding 0.95).
- By comparing pairs of highly correlated features, we retained the one that had a stronger linear relationship with the target (bad_flag) and dropped the other, significantly reducing dimensionality.





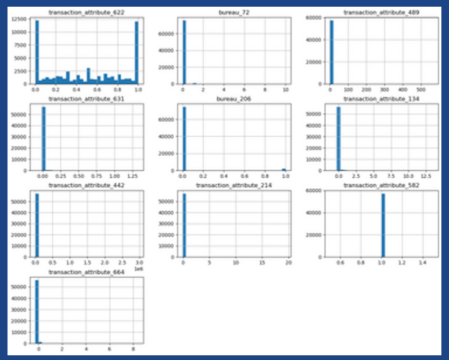
Domain Insight

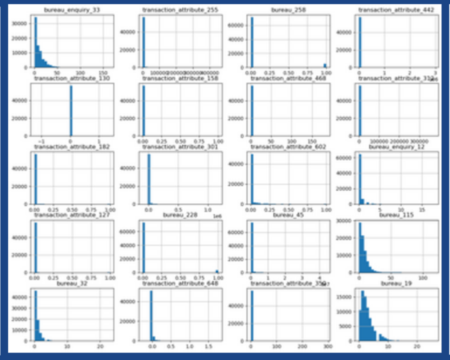


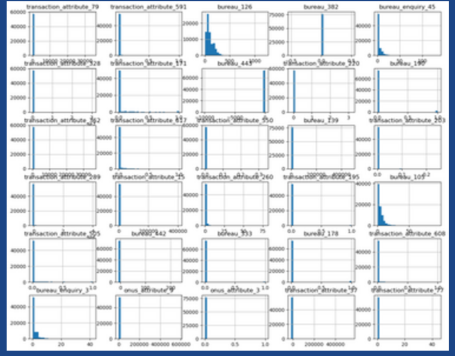


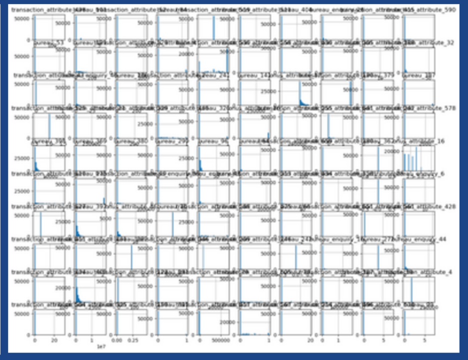
- “onus_attribute_1” we observed that >99.9% defaulters were under the credit-limit of ~500k (Rs?)
- Almost all non-default accounts have onus_attribute_26 ≈ 0.
- Almost all default accounts have onus_attribute_26 ≈ 1.
- onus_attribute_1 - Credit Limit or Monthly ATM txns
- onus_attribute_26 - Home Ownership or Employment Verfication

Trend Suggested vast number of features did not correlate with the target by our estimate





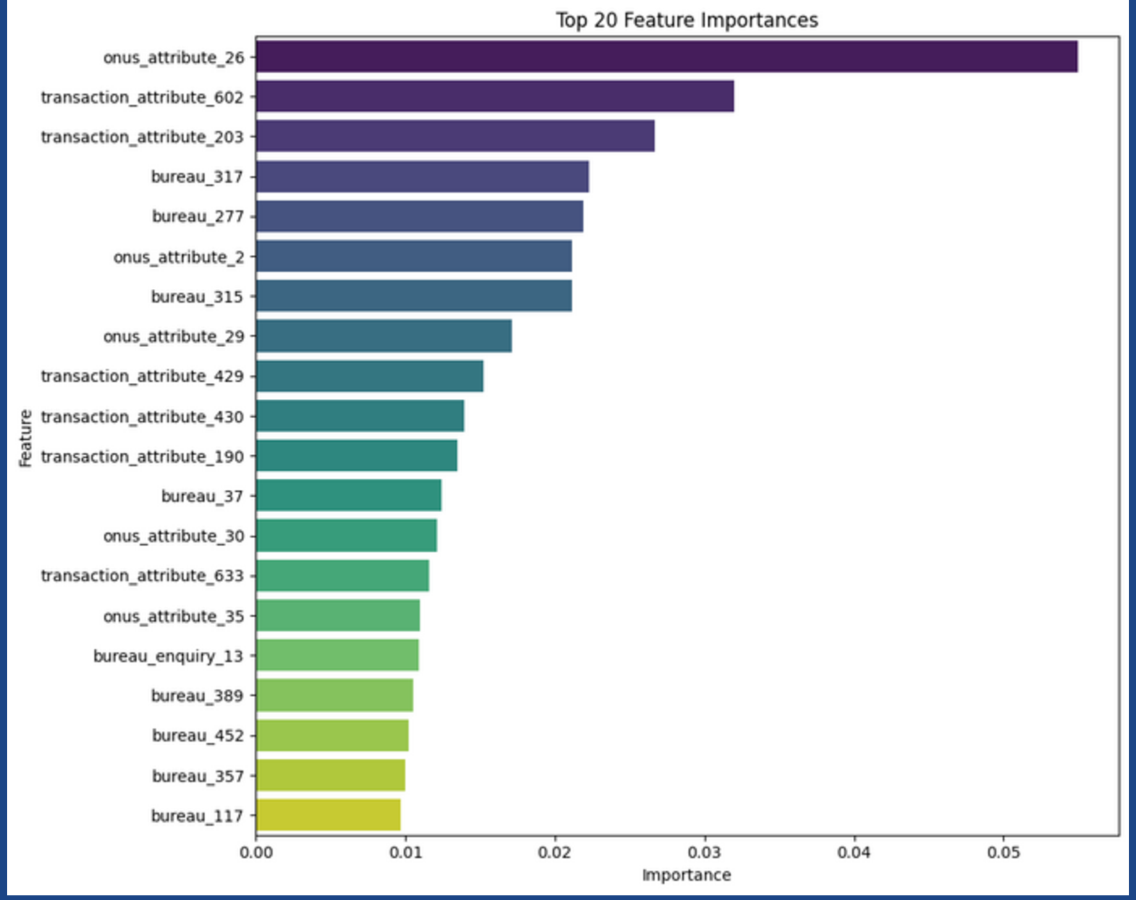




FEATURE IMPORTANCE AND SELECTION

Feature Importance

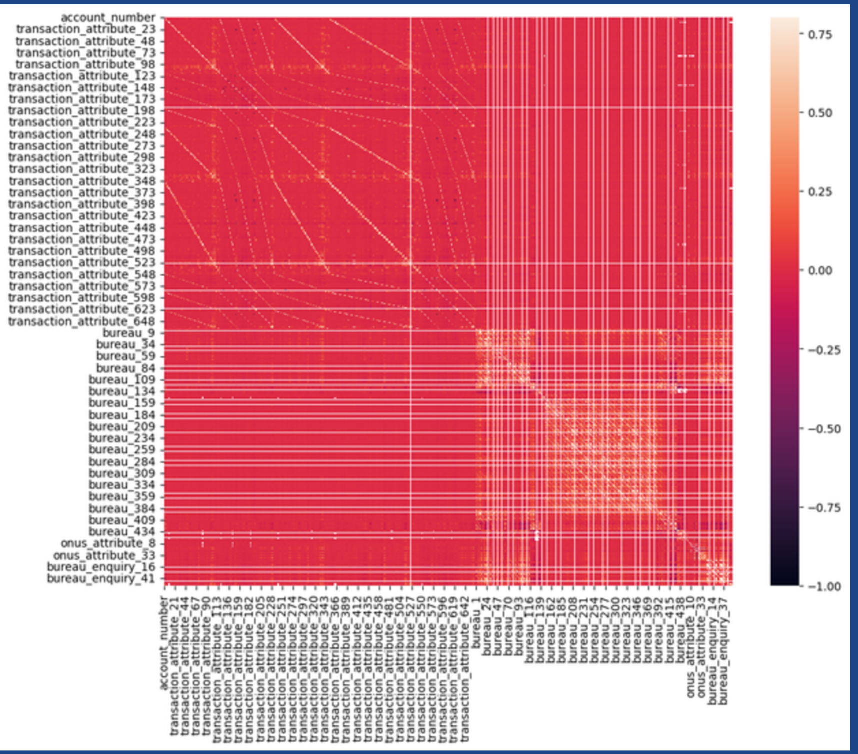
- onus_attribute_26
- transaction_attribute_602
- transaction_attribute_203
- bureau_317
- bureau_277



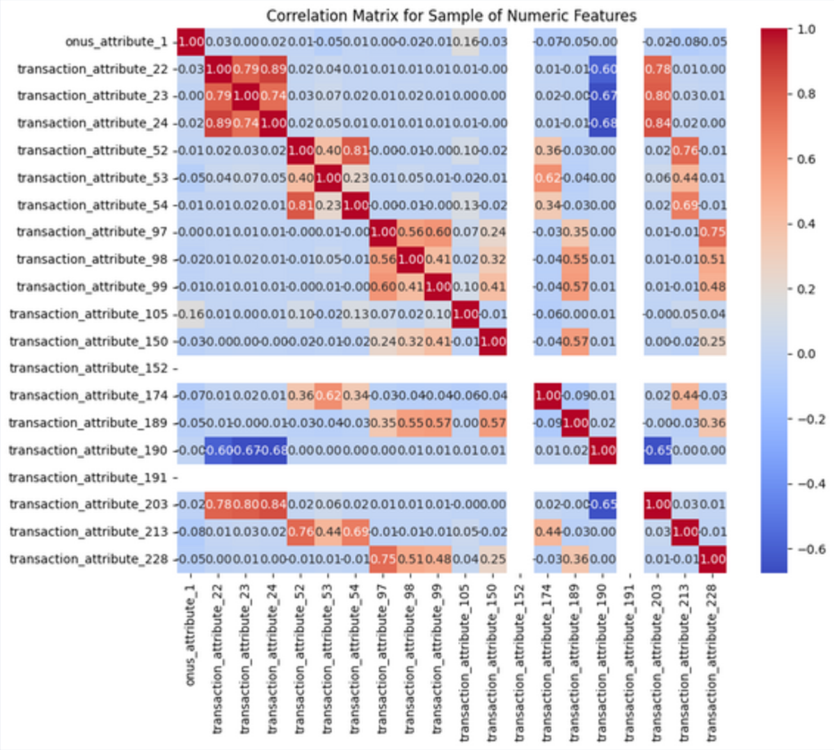
Dimensionality reduction - Statistical Techniques

- Identified highly correlated data pairs, removed the feature less correlated with target.
- Identified that features were linearly correlated by Pearson-Spearman Difference.
- Since linearity was important, features with less that 0.065*0.3 = 0.017 correlation index were discarded.
- Final Dimensions went from 77,444 by 1216 to 61955 by 257.

BEFORE



AFTER

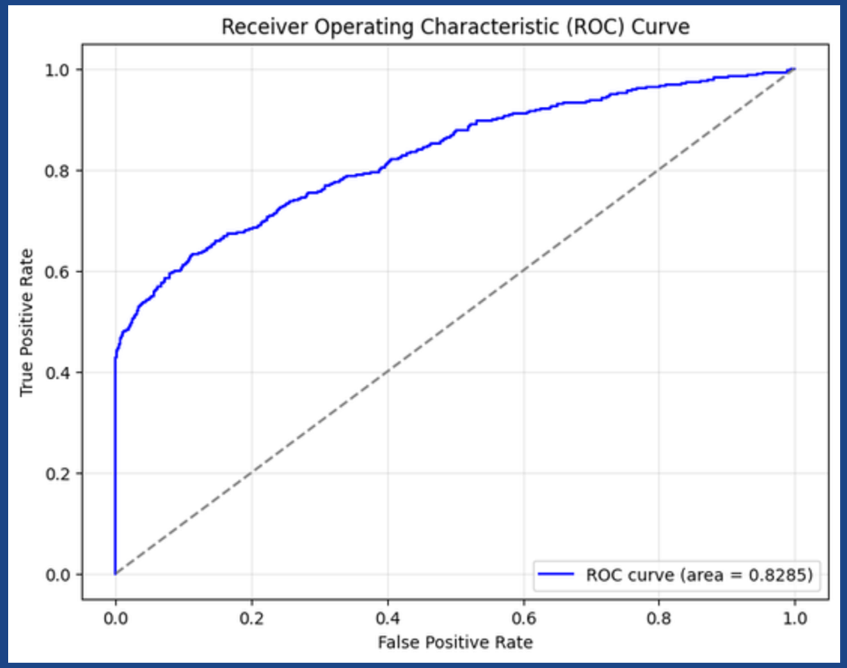


MODEL EVALUATION AND METRICS

- Performance comparison table

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.9839	0.1081	0.0182	0.0311	0.7599
XGBoost (Default)	0.9866	0.25	0.0049	0.0096	0.7673
Trained XGBoost	0.9869	0.5545	0.4553	0.5	0.8285

- Accuracy was a bad metric for such unbalanced data since ~98.6% dataset is Non-defaulting.
- Precision: Ensures flagged fraud cases are truly fraudulent
- Recall: Captures as many fraud cases as possible
- F1 Score: Balances precision and recall for a better trade-off.
- ROC-AUC: Evaluates model performance across different classification thresholds.
- Confusion Matrix: Visualizes false positives & false negatives for better understanding




```
Confusion Matrix:
[[34274   184]
 [  274   229]]
```

FINAL MODEL INSIGHTS

Why the Final Model Worked Best

Rigorous Data Tailoring:

• Removed unreliable and redundant features (high missingness, high inter-feature correlation).
• Focused on features with at least minimal predictive signal (dropped near-zero correlation features).

Handling Imbalance:

• Applied stratified splitting and class reweighting to ensure the model focused on the minority class (defaulters).

Capturing Non-linear Relationships:

• Identified non-linear dependencies using mutual information—allowing flexible models (e.g., tree-based/ensemble) to capture complex patterns.

Meaning of Prediction Insights

Actionable Signals:

• Features like “onus_attribute_26” provide clear thresholds that distinguish defaulters from non-defaulters.
• Key indicators reflect internal behaviours (e.g., on-us transactions) that are closely tied to credit risk.

Robust Discrimination:

• A high ROC AUC (>0.83) demonstrates strong ranking ability even though overall accuracy is inflated by class imbalance. • Focus on recall ensured that critical defaulters are less likely to be missed.

Explainability

Transparent Feature Engineering:

• The model’s inputs were carefully selected based on statistical and domain insights, making it easier to justify why certain features matter.

Interpretable Signals:

• Binary indicators and threshold-based features allow stakeholders to understand decision rules (e.g., nearly all defaulters have onus_attribute_26 = 1).

Model Output:

• By emphasizing feature importance and using visualization (density/strip plots), the model’s predictions can be traced back to clear, actionable factors.

THANK YOU