Credit Data Analyst – Final Report

1. Synthetic Data Creation

1.1. Features Included & Rationale

Data Module	Data Field	Business Description	Rationale	Data Source
Business Profile	business_size	Number of employees (e.g., small, medium, large).	Larger businesses are generally lower risk.	Business registration
	business_time	Number of years the business has been operating.	Older businesses are typically more stable.	
	business_industry	Industry sector (e.g., retail, hospitality, manufacturing).	Industry-specific risks (e.g., hospitality is seasonal).	
	business_structure	Legal structure (e.g., sole trader, partnership, company).	Different structures have varying risk profiles.	
	business_location	Geographic location (e.g., urban, suburban, rural).	Location impacts market size and economic conditions.	
	business_job_openings	Job Openings in career websites (Seek, Indeed, Linkedin)	Indicates business growth or hiring needs	3rd party platform
	business_review	Average Google review score (e.g., 1-5 stars).	Reflects customer satisfaction and reputation.	
Platform Engagement	plat_login_cnt_t1m	Number of logins to Zeller's platform in the last 1 month.	Indicates engagement with Zeller's services.	Zeller platform/App
	plat_customer_support_cnt_t3m	Number of customer support interactions in the last 3 months.	Reflects the business's reliance on support.	
	plat_dashboard_cnt_t3m	Number of Zeller Dashboard usage in the last 3 months.	Indicates proactive financial management.	
	plat_terminal_flag	Whether the business owns a Zeller terminal.	Ownership of a terminal may indicate availability of	
	plat_trsct_acct_flag	Whether the business owns a Zeller transaction account.	Indicates reliance on Zeller for core financial operations.	
	plat_saving_acct_flag	Whether the business owns a Zeller savings account	Indicates reliance on Zeller for core financial operations.	
Finance	fin_revenue	Total revenue last finiancial year	Indicates the business's ability to generate income.	Financial statements,
	fin_profit	Net profit last finiancial year	Measures financial efficiency and profitability.	self-report
	fin_assets	Total assets last finiancial year	Indicates the business's financial strength and resource base.	
	fin_liailities	Total liabilities last financial year	Indicates the business's .	
Transactions	tran_volume_t1m	Total transaction volume in the last 1 month.	Reflects recent business activity.	Zeller payment portal
	tran_cnt_t1m	Total number of transactions over the last 1 month.	Reflects recemt business activity.	
	tran_volume_t3m	Total transaction volume over the last 3 months.	Reflects mid-term business activity.	
	tran_cnt_t3m	Total number of transactions over the last 3 months.	Reflects mid-term business activity.	
	tran_inflow_avg_t3m	Average cash inflow over the last 3 months.	Indicates the business's ability to generate cash.	
	tran_outflow_avg_t3m	Average cash outflow over the last 3 months.	Measures the business's spending patterns.	
Credit History	credit_score	Business credit score (e.g., 0-1000).	A direct measure of creditworthiness.	Credit Bureau such as
	credit_lines_cnt	Number of active credit.	Indicates the business's reliance on credit.	Equifax, Experian, illion
	credit_current_amt	Total outstanding credit amount.	Measures the business's debt burden.	
	credit_default_cnt_t12m	Number of defaults in the last 12 months.	Predicts future default risk.	
	credit_inquiry_cnt_t3m	Number of credit inquiries in the last 3 months.	Reflects short-term credit-seeking behavior.	
	credit_inquiry_cnt_t12m	Number of credit inquiries in the last 12 months.	Reflects long-term credit-seeking behavior.	
	credit_court_cnt_t12m	Number of court judgments against the business.	Indicates legal or financial issues.	
Loan	loan_amount	Total loan amount outstanding.	Measures the business's debt burden.	Zeller loan records
	loan_utilization_amount	Proportion of available credit being used.	High utilization may indicate over-leverage.	
	loan_late_repayment_cnt	Number of Late repayments within 90 days	Indicates repayment behavior and potential default risk.	
	loan_start_date	Date of origination of the loan	Helps assess the loan's age and repayment progress.	
	loan_maturity_date	Date of maturity of the loan	Helps assess the loan's tenure and uncertainty.	
	loan_interest_rate	Interest rate applied to the loan	Higher interest rate increase repayment burden	
Default	default	Default or not for the loan in assessment	Used as Y to predict future default risk.	

Appendix 1

1.2. Assumptions

a. Definition of Default (Assume at 8% but resampled to 20%):

a) **90 days past due** on any scheduled payment.
b) Material breaches of loan terms (e.g., financial covenants).
c) Bankruptcy or certain legal actions as definitive indicators of default.

b. Reference Date (2024-10-01)

a) allows sufficient time to observe payment behavior and identify defaults accurately.
b) provides a 3-month window (2024-10-01 to 2024-12-31) for gathering validation data for PSI test
c) aligns with practical timelines for data collection, model validation, and deployment, ensuring the project remains on schedule.

c. Data Access:

a) Zeller has registration with credit bureaus (e.g., Equifax, Experian, illion) to retrieve credit score and related data.

b) Zeller has access to Google/Yelp APIs to retrieve business review data,

d. Regulatory Compliance:

a) Zeller adheres to all relevant regulations (e.g., Privacy Act 1988, Australian Privacy Principles) when collecting, storing, and using data.
b) Customers are willing to provide financial statements and consent for data sharing.

e. Data Generation:

a) Ranges and distributions for synthetic data are based on personal understanding and certain assumptions on values based on proportion are made to be more realistic (e.g. 80% are not default and 80% of business have no court judgements, etc.)
b) limited time spans are used to simplify the process. (In real cases, hundreds of variables can be fed into the algorithm for more accurate risk assessment.)
c) In real-world, data need to be cleaned to handling missing values, outliers, mistakes,
d) In real-word, data are gathered from different sources. ETL process may be needed to get the flat table with all original features.

1.3. Limitations on Features Choice & Data Creation

a. Data Sources

a) Credit Data: Subject to client's consent for access, and may not reflect recent financial changes.
b) Financial Data: Relies on customer self-disclosure, which may not always be accurate, complete, or up-to-date.
c) Business Registration Data: May not be updated in real-time, leading to outdated or incomplete information.
d) Transaction Data: Requires customer consent for Zeller's usage, and may be limited to transactions processed through Zeller's platform. There may also be fraud data that Zeller POS should identify.
e) Platform Behavior Data: Dependent on the availability of data tracking

b. Data Quality:

a) Synthetic or auto-generated data may lack real-world patterns, reducing the accuracy of models.
b) Some variables may have real-life boundaries on others, e.g. large business may have higher revenue, auto-generated data cannot realize all hidden patterns.

c. Data Sampling:

a) 10,000 records may be inadequate for meaningful patterns that can be generalized
b) data is resampled to relieve the imbalance of label and produce more reliable evaluation and feature importance analysis, however, this may introduce bias in model predictions, lead to overfitting of resampled data. To mitigate this, we can adjust the threshold to make final default or not decision based on real-word default rate and business objectives.

2. Feature Engineering

2.1. Feature Derivation

Common used techniques include as below, however, due to limit of step 1, some data are purely generated here, but will also showcase calculations given full original data provided.

2.1.1. Frequency

Definition: Count of occurrences of a specific event or behavior
Example - Existing: credit_inquiry_cnt_t3m

2.1.2. Sum

Definition: Amount of specific value over a period
Example - Existing: fin_revenue

2.1.3. Average

Definition: Mean value of a specific metric over a period
Example - Derived:

- drv_profit_pm_avg_t12m (net_profit/12)
- drv_income_pm_avg_t3m (inflow-outflow/3)
- drv_tran_volume_pm_avg_t3m
- drv_tran_cnt_pm_avg_t3m

2.1.4. Ratio

Definition: Ratio or proportion of one metric relative to another.
Example - Derived:

- drv_profit_margin_rate (profit/revenue)
- drv_asset_liability_rate (asset/liability)
- drv_loan_util_rate (utilization amount/loan amount)
- drv_inflow_outflow_rate_t3m (inflow/outflow)

2.1.5. Consistency

Definition: Measures the stability or variability of a metric over time.
Example - Not available here due to lack of data:

- drv_addr_cons_flag (whether google map location the same with business registration)
- drv_industry_cons_flag

2.1.6. Recency

Definition: Measures how recent a specific event or behavior occurred.
Example - provided: business_time
Example - derived:

- drv_mon_since_apply
- drv_mon_since_last_late (the last late data is not provided in step1, so this would be purely generated)

2.1.7. Standard Deviation

Definition: Measures the variability or dispersion of a metric.
Example - derived:

- drv_tran_volume_std_t12m: Standard deviation of transaction volume over the last 6 months.
- drv_inflow_std_t6m:Standard deviation of cash inflow over the last 6 months.

2.1.8. Proportion

Definition: Ratio of a subset to the total.
Example - derived:

- drv_trans_vol_prop_t1t3
- drv_credit_inquiry_prop_t3t12

2.2. Feature Transformation

Feature selection is essential to reduce dimensionality, improve model performance, and avoid overfitting. Different techniques were applied based on the model type.

2.2.1. Logistic Regression (LR)

For Logistic Regression, Weight of Evidence (WoE) transformation was applied to the features. WoE is a technique used to transform categorical and continuous variables into a format that captures the relationship between the feature and the target variable (default). This transformation helps in:

- Linearizing the relationship between features and the target.
- Handling missing values by binning them into a separate category.
- Reducing the impact of outliers by grouping extreme values. The WoE transformation was performed using the scorecardpy.woebin function, which automatically bins the features based on their relationship with the target. Manual adjustments were made to the binning process for specific features (e.g., drv_loan_util_rate and fin_revenue) to ensure meaningful groupings.

2.2.2. XGBoost (XGB)

For XGBoost, one-hot encoding was applied to categorical features to convert them into a binary format. This ensures that the model can interpret categorical variables effectively. Additionally, missing values were filled with -99 to handle any null values in the dataset.

2.2.3. Artificial Neural Network (ANN)

For ANN, one-hot encoding was also applied to categorical features. Additionally, standard scaling was performed to normalize the features. ANN models are sensitive to the scale of input features, and scaling ensures that all features contribute equally to the model's learning process.

2.3. Feature Selection

Three models were trained: Logistic Regression (LR), XGBoost (XGB), and Artificial Neural Network (ANN). Hyperparameter tuning was performed for each model to optimize performance.

2.3.1. Logistic Regression (LR)

Feature selection for LR was performed in two steps:

a. Initial Selection Based on Information Value (IV):

Features with an IV below 0.02 were removed, as they provide little predictive power.
Key features like plat_login_cnt_t1m and business_time were explicitly retained.

b. Stepwise Selection Based on p-values:

A forward-backward stepwise selection process was used to select features based on their statistical significance (p-value < 0.05 for inclusion, p-value > 0.1 for exclusion).
This process ensures that only the most relevant features are included in the final model.

2.3.2. XGBoost (XGB)

For XGBoost, Recursive Feature Elimination (RFE) was used to select the top 40 features based on their importance. RFE recursively removes the least important features and ranks the remaining features based on their contribution to the model's performance.

2.3.3. Artificial Neural Network (ANN)

For ANN, correlation-based filtering was first applied to remove highly correlated features (correlation > 0.8). This reduces redundancy and ensures that the model does not overfit. After this, RFE was used to select the top 40 features based on their importance.

3. Model Training

3.1. Logistic Regression (LR)

3.1.1. Rationale

- Regulatory Compliance: LR is often preferred in industries like banking and finance because it aligns with regulatory requirements (e.g., Basel III, IFRS 9) that emphasize model transparency.
- Feature Importance: Techniques like Information Value (IV) and Weight of Evidence (WoE) make it easy to understand the contribution of each feature to the model's predictions.
- Stability: LR is less prone to overfitting, especially when combined with feature selection techniques like stepwise regression.

3.1.2. Pros

- Interpretability: Coefficients provide clear insights into the relationship between features and the target variable.
- Efficiency: LR is computationally inexpensive and easy to implement.
- Regulatory Acceptance: Its simplicity makes it easier to explain to regulators and stakeholders.
- Robustness: Performs well even with smaller datasets.

3.1.3. Cons

- Linearity Assumption: LR assumes a linear relationship between features and the log-odds of the target, which may not capture complex patterns.
- Feature Engineering: Requires careful feature engineering (e.g., WoE binning) to handle non-linear relationships.
- Limited Flexibility: May underperform compared to more complex models when dealing with highly non-linear data.

3.1.4. Hyperparameters

- penalty: Regularization type (l1, l2, elasticnet).
- C: Inverse of regularization strength (values: 0.5, 0.6, 0.7, 0.9, 1.0).
- solver: Optimization algorithm (liblinear, saga).

Rationale:

- Regularization helps prevent overfitting by penalizing large coefficients.
- C controls the trade-off between fitting the training data and keeping the model simple.
- The solver was chosen based on compatibility with the regularization type.

3.2. XGBoost (XGB)

3.2.1. Rationale

- Competitions and Real-World Applications: XGBoost has been a top performer in machine learning competitions (e.g., Kaggle) and is widely adopted in industries like fintech and banking.
- Feature Importance: XGBoost provides built-in feature importance metrics, which help identify key drivers of credit risk.
- Handling Imbalanced Data: Techniques like class weighting and sampling make XGBoost effective for imbalanced datasets (e.g., low default rates

3.2.2. Pros

- High Accuracy: XGBoost often outperforms traditional models like LR in predictive performance.
- Flexibility: Can handle both numerical and categorical features without extensive preprocessing.
- Scalability: Efficiently handles large datasets and high-dimensional feature spaces.
- Regularization: Built-in regularization techniques (e.g., L1, L2) reduce overfitting.

3.2.3. Cons

- Complexity: Less interpretable than LR, making it harder to explain to regulators.
- Computational Cost: Training can be resource-intensive, especially for large datasets.
- Hyperparameter Tuning: Requires careful tuning of hyperparameters (e.g., learning rate, max depth) to achieve optimal performance.

3.2.4. Hyperparameters

- learning_rate: Step size shrinkage (values: 0.01, 0.5, 0.1, 0.2).
- max_depth: Maximum depth of the tree (values: 3, 4, 5, 6, 7).

Rationale:

- learning_rate controls the contribution of each tree to the final model.
- max_depth controls the complexity of the model by limiting the depth of the trees.

3.3. Artificial Neural Network (ANN)

3.3.1. Rationale

- Deep Learning Adoption: ANN and deep learning models are increasingly used in industries like fintech for tasks like fraud detection and credit scoring.
- Feature Learning: ANN can automatically learn relevant features from raw data, reducing the need for manual feature engineering.
- Scalability: ANN can handle large datasets and high-dimensional feature spaces effectively.

3.3.2. Pros

- High Flexibility: ANN can model complex, non-linear relationships that simpler models cannot capture.
- Feature Learning: Automatically learns relevant features, reducing the need for manual feature engineering.
- Scalability: Performs well on large datasets and high-dimensional feature spaces.
- State-of-the-Art Performance: Often achieves the highest predictive accuracy among all models.

3.3.3. Cons

- Black-Box Nature: ANN is highly complex and difficult to interpret, making it less suitable for regulatory compliance.
- Computational Cost: Training ANN is resource-intensive and requires significant computational power.

- Data Requirements: ANN requires large amounts of data to avoid overfitting.
- Hyperparameter Tuning: Requires careful tuning of hyperparameters (e.g., learning rate, hidden layers) to achieve optimal performance.

3.3.4. Hyperparameters

- hidden_layer_sizes: Number of neurons in each hidden layer (values: (50,), (100,), (100, 50)).
- activation: Activation function (relu, sigmoid, tanh).
- solver: Optimization algorithm (adam, sgd).
- alpha: L2 regularization term (values: 0.0001, 0.001).

Rationale:

- hidden_layer_sizes determines the capacity of the network to learn complex patterns.
- The activation function introduces non-linearity into the model.
- alpha helps prevent overfitting by penalizing large weights.

3.4. Model Comparison

- Start with LR: Use Logistic Regression for baseline modeling and regulatory compliance.
- Adopt XGBoost: Use XGBoost for high-performance requirements and complex datasets.
- Explore ANN: Use ANN for large-scale datasets and advanced modeling tasks where interpretability is not a primary concern.

Model	Interpretability	Performance	Scalability	Regulatory Compliance	Use Case
LR	High	Moderate	High	High	Baseline models, regulatory reporting
XGBoost	Moderate	High	High	Moderate	High-performance, complex datasets
ANN	Low	Very High	High	Low	Large-scale, advanced modeling tasks

4. Model Evaluation

4.1. Metrics Selection

UC and KS are widely used in credit risk modeling because they provide complementary insights into the model's performance. AUC measures the overall ability to distinguish between classes, while KS focuses on the separation of classes at specific thresholds. Together, they help ensure that the model is both accurate and practical for real-world decision-making.

4.1.1. AUC (Area Under the Curve)

Definition:

AUC stands for Area Under the Receiver Operating Characteristic (ROC) Curve. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

- True Positive Rate (TPR): The proportion of actual positives correctly identified by the model (also called sensitivity or recall).
- False Positive Rate (FPR): The proportion of actual negatives incorrectly identified as positives by the model.

Interpretation:

AUC measures the model's ability to distinguish between positive (default) and negative (non-default) classes. It ranges from 0 to 1:

- AUC = 0.5: The model performs no better than random guessing.

- AUC = 1: The model perfectly distinguishes between the two classes.
- AUC > 0.7: Generally considered acceptable, with higher values indicating better performance.

Rationale:

- Discriminatory Power: AUC measures how well the model can distinguish between good (non-default) and bad (default) customers. This is critical in credit risk modeling, where the goal is to accurately identify high-risk customers.
- Threshold Independence: AUC is independent of the classification threshold, making it a robust metric for evaluating model performance across different threshold settings.
- Interpretability: AUC provides a single, intuitive metric that summarizes the model's performance across all possible thresholds.

4.1.2. KS

Definition:

The KS statistic measures the maximum difference between the cumulative distribution functions (CDFs) of the positive and negative classes. It is calculated as:

$$KS = \max\left(|F_{positive}(x) - F_{negative}(x)|\right)$$

where $F_{positive}(x)$ and $F_{negative}(x)$ are the CDFs of the predicted probabilities for the positive and negative classes, respectively.

Interpretation:

KS ranges from 0 to 1:

- KS = 0: The model cannot distinguish between the two classes.
- KS = 1: The model perfectly separates the two classes.
- KS > 0.3: Generally considered acceptable, with higher values indicating better separation.

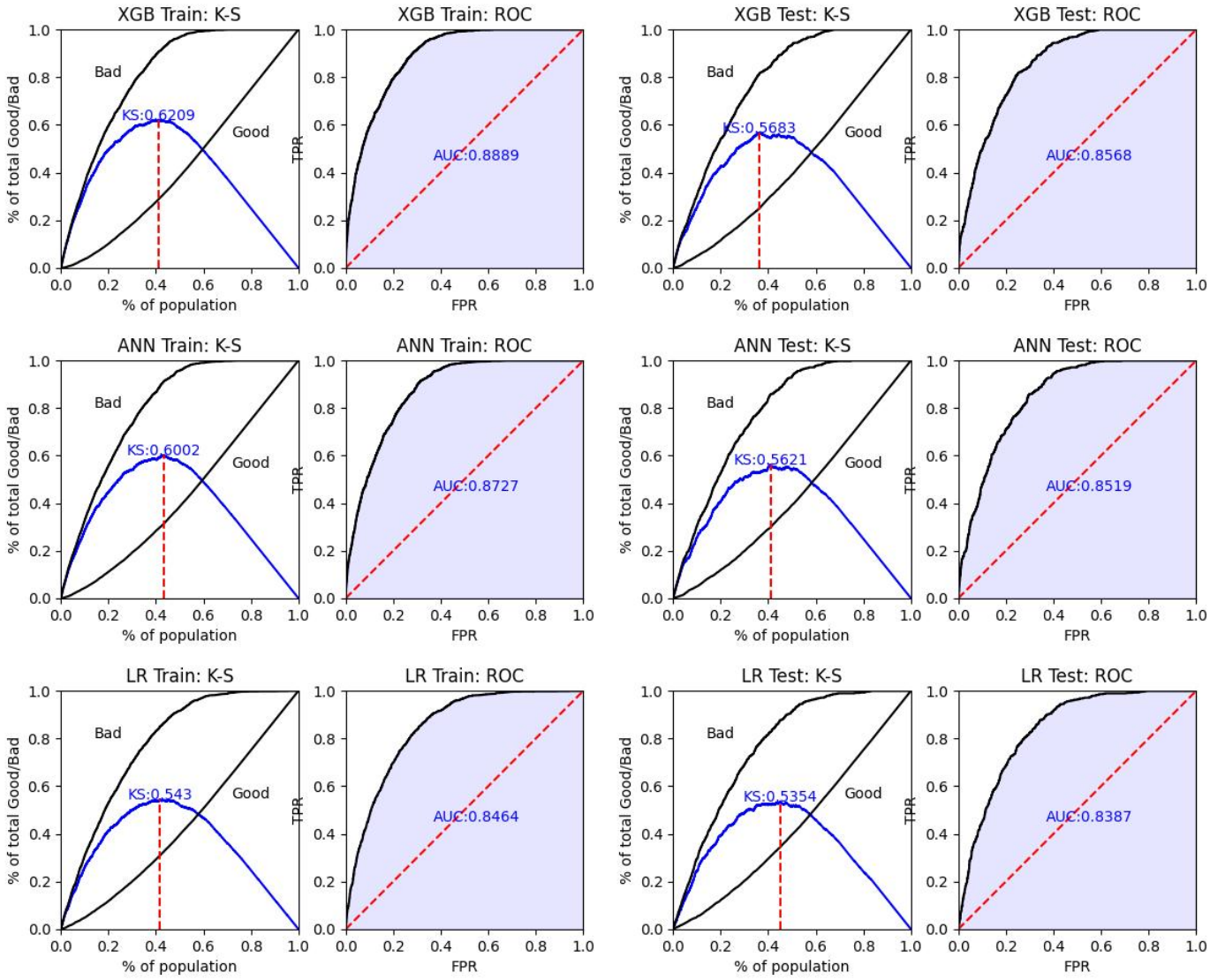
Rationale

- Separation of Classes: KS directly measures the separation between the distributions of predicted probabilities for the positive and negative classes. A higher KS indicates better separation, which is essential for effective risk scoring.
- Threshold Selection: KS helps identify the optimal threshold for classification by highlighting the point where the difference between the two distributions is maximized.
- Model Calibration: KS is useful for assessing whether the model's predicted probabilities are well-calibrated, i.e., whether the probabilities reflect the true likelihood of default.

4.2. Model Performance

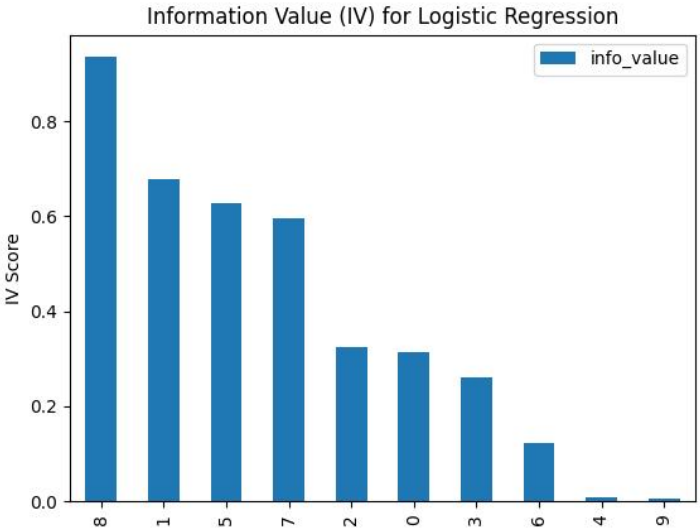
Model	Train AUC	Test AUC	Train KS	Test KS
LR	0.8432	0.8435	0.5414	0.5392
ANN	0.8787	0.8584	0.6018	0.5617
XGB	**0.8859**	**0.8621**	**0.6141**	**0.5788**

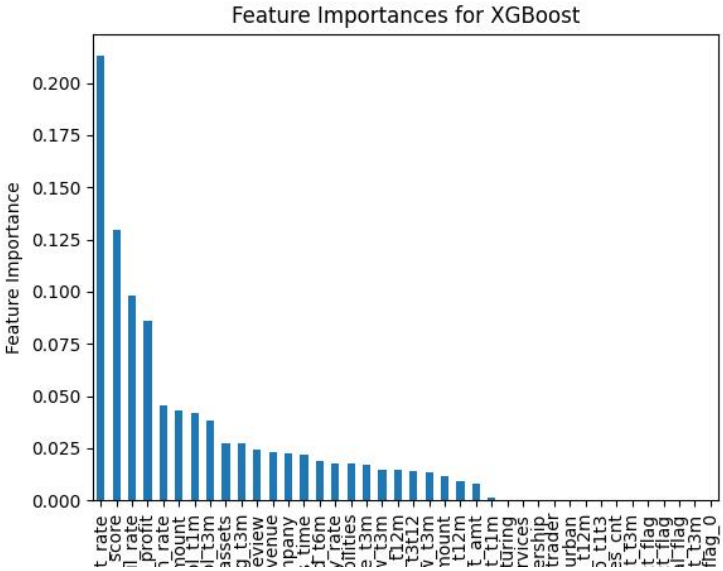
in Plotting:

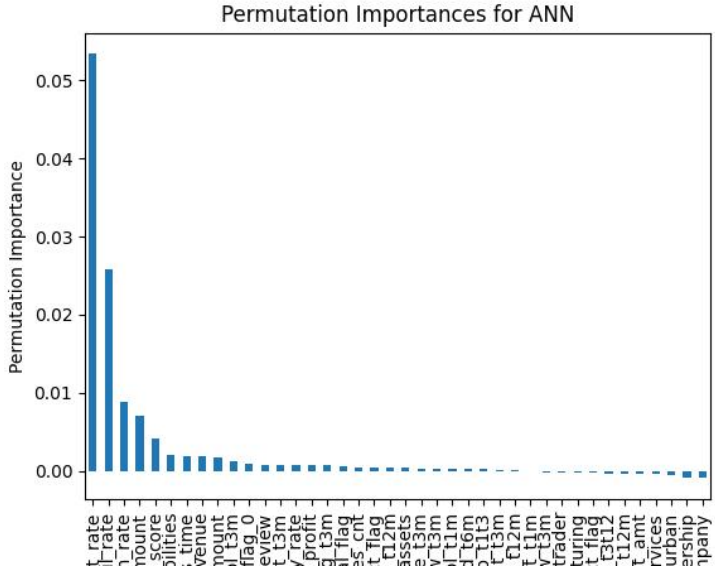


4.3. Feature Importance

refer to Appendix 2







4.4. Key Observations

- XGBoost and ANN outperformed Logistic Regression in terms of both AUC and KS, but Logistic Regression is the most stable one.
- The performance gap between train and test metrics in XGBoost and ANN was minimal, indicating that the models might be overfit.
- XGBoost achieved the highest AUC and KS, suggesting it is the best-performing model.
- Logistic Regression is the most interpretable model given the binning image

5. Insights

5.1. Performance Metrics

- LR's simplicity and interpretability make it a reliable choice for regulatory compliance and stakeholder communication.
- XGBoost's is the best-performed model within this project, its ability to handle non-linear relationships and complex datasets makes it a powerful tool for high-performance credit risk modeling.
- ANN's ability to capture intricate patterns in data makes it suitable for large-scale datasets, though its

black-box nature limits interpretability.

5.2. Key Drivers of Credit Risk

All results are based on auto-generated synthetic data. In real-world scenarios, the performance and feature importance may differ due to variations in data quality, distribution, and domain-specific nuances.

5.2.1. LR

loan_interest_rate_woe, credit_score_woe, and fin_profit_woe are the most influential features, highlighting the importance of financial health and credit history in predicting default.

5.2.2. XGBoost

loan_interest_rate, credit_score, and drv_loan_util_rate are the top contributors, which is consistent with findings from LR and also emphasizes the role of loan utilization in credit risk.

5.2.3. ANN

loan_interest_rate, drv_loan_util_rate, and credit_score are the most important, consistent with the findings from LR and XGBoost.

5.3. Actionable Insights (based on key drives about)

6. Future Enhancements

6.1. Real-World Data

Replace synthetic data with actual credit data to ensure the models are trained on realistic and domain-specific datasets. This will improve the generalizability and reliability of the models in real-world applications.

6.2. Model Explainability

ANN and XGBoost can be challenging to interpret directly. Enhance explainability using techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to provide insights into model predictions and ensure compliance with regulatory requirements.

6.3. Model Stability

Use the Population Stability Index (PSI) to measure the distribution shift between training data and future data. Validate model stability using data from 2024-10-01 to 2024-12-31:
PSI < 0.1: Excellent stability.
0.1 < PSI < 0.2: Acceptable stability.
PSI > 0.2: Unstable model; requires investigation and adjustment.

6.4. Deployment

Integrate the models into a production environment for real-time risk assessment. Ensure seamless integration with existing systems and workflows to enable timely and accurate decision-making.

6.5. Consider AutoML

Automated Machine Learning (AutoML) can quickly generate baseline models for comparison with custom-built models. However, in credit risk modeling, interpretability is often critical for regulatory compliance and stakeholder trust. A hybrid approach—using AutoML for rapid prototyping and refining models manually—can balance efficiency and compliance.

Appendix 1

Data Module	Data Field	Business Description	Rationale	Data Source
Business Profile	business_size	Number of employees (e.g., small, medium, large).	Larger businesses are generally lower risk.	Business registration
	business_time	Number of years the business has been operating.	Older businesses are typically more stable.	
	business_industry	Industry sector (e.g., retail, hospitality, manufacturing).	Industry-specific risks (e.g., hospitality is seasonal).	
	business_structure	Legal structure (e.g., sole trader, partnership, company).	Different structures have varying risk profiles.	
	business_location	Geographic location (e.g., urban, suburban, rural).	Location impacts market size and economic conditions.	
	business_job_openings	Job Openings in career websites (Seek, Indeed, Linkedin)	Indicates business growth or hiring needs	3rd party platform
	business_review	Average Google review score (e.g., 1-5 stars).	Reflects customer satisfaction and reputation.	
Platform Engagement	plat_login_cnt_t1m	Number of logins to Zeller's platform in the last 1 month.	Indicates engagement with Zeller's services.	Zeller platform/App
	plat_customer_support_cnt_t3m	Number of customer support interactions in the last 3 months.	Reflects the business's reliance on support.	
	plat_dashboard_cnt_t3m	Number of Zeller Dashboard usage in the last 3 months.	Indicates proactive financial management.	
	plat_terminal_flag	Whether the business owns a Zeller terminal.	Indicate availability of transaction data	
	plat_trsct_acct_flag	Whether the business owns a Zeller transaction account.	Indicates reliance on Zeller for core financial operations.	
	plat_saving_acct_flag	Whether the business owns a Zeller savings account	Indicates reliance on Zeller for core financial operations.	
Finance	fin_revenue	Total revenue last financial year	Indicates the business's ability to generate income.	Financial statements, self-report
	fin_profit	Net profit in last financial year	Measures financial efficiency and profitability.	
	fin_assets	Total assets last financial year	Indicates the business's financial strength and resource base.	
	fin_liabilities	Total liabilities last financial year	Indicates the business's .	
Transactions	tran_vol_t1m	Total transaction volume in the last 1 month.	Reflects recent business activity.	Zeller payment portal
	tran_cnt_t1m	Total number of transactions over the last 1 month.	Reflects recemt business activity.	
	tran_vol_t3m	Total transaction volume over the last 3 months.	Reflects mid-term business activity.	
	tran_cnt_t3m	Total number of transactions over the last 3 months.	Reflects mid-term business activity.	
	tran_inflow_t3m	Total cash inflow over the last 3 months.	Indicates the business's ability to generate cash.	

	tran_outflow_t3m	Total cash outflow over the last 3 months.	Measures the business's spending patterns.	
Credit History	credit_score	Business credit score (e.g., 0-1000).	A direct measure of creditworthiness.	Credit Bureau such as Equifax, Experian, illion
	credit_lines_cnt	Number of active credit.	Indicates the business's reliance on credit.	
	credit_current_amt	Total outstanding credit amount.	Measures the business's debt burden.	
	credit_default_cnt_t12m	Number of defaults in the last 12 months.	Predicts future default risk.	
	credit_inquiry_cnt_t3m	Number of credit inquiries in the last 3 months.	Reflects short-term credit-seeking behavior.	
	credit_inquiry_cnt_t12m	Number of credit inquiries in the last 12 months.	Reflects long-term credit-seeking behavior.	
	credit_court_cnt_t12m	Number of court judgments against the business.	Indicates legal or financial issues.	
Loan	loan_amount	Total loan amount outstanding.	Measures the business's debt burden.	Zeller loan records
	loan_utilization_amount	Proportion of available credit being used.	High utilization may indicate over-leverage.	
	loan_late_repayment_cnt	Number of Late repayments within 90 days	Indicates repayment behavior and potential default risk.	
	loan_start_date	Date of origination of the loan	Helps assess the loan's age and repayment progress.	
	loan_maturity_date	Date of maturity of the loan	Helps assess the loan's tenure and uncertainty.	
	loan_interest_rate	Interest rate applied to the loan	Higher interest rate increase repayment burden	
Default	default	Default or not for the loan in assessment	Used as Y to predict future default risk.	

Appendix 2

Model	Metrics	variable	Value
LR	info_value	loan_interest_rate_woe	0.9351
		credit_score_woe	0.6769
		fin_profit_woe	0.6287
		drv_profit_pm_avg_t12m_woe	0.5955
		business_size_woe	0.3244
		loan_utilization_amount_woe	0.2607
		plat_login_cnt_t1m_woe	0.0088
ANN	permutation_importance	loan_interest_rate	0.0475
		drv_loan_util_rate	0.0281
		credit_score	0.0056

		fin_profit	0.0046
		loan_utilization_amount	0.0039
		drv_profit_margin_rate	0.0033
		plat_customer_support_cnt_t3m	0.0023
		fin_assets	0.0021
		business_structure_partnership	0.0020
		drv_inflow_outflow_rate_t3m	0.0019
		credit_lines_cnt	0.0018
		business_industry_retail	0.0016
		tran_outflow_t3m	0.0016
		credit_court_cnt_t12m	0.0016
		drv_income_pm_avg_t3m	0.0016
		business_location_suburban	0.0014
		plat_dashboard_cnt_t3m	0.0012
		drv_inflow_std_t6m	0.0012
		drv_asset_liability_rate	0.0011
		tran_inflow_t3m	0.0011
		drv_trans_vol_prop_t1t3	0.0011
		drv_tran_vol_std_t12m	0.0011
		fin_liabilities	0.0011
		business_review	0.0011
		tran_cnt_t3m	0.0010
XGB	feature_importance	loan_interest_rate	0.2092
		credit_score	0.1114
		drv_loan_util_rate	0.0966
		fin_profit	0.0962
		loan_utilization_amount	0.0484
		drv_profit_margin_rate	0.0431
		tran_cnt_t3m	0.0283
		tran_vol_t1m	0.0269
		tran_vol_t3m	0.0211
		drv_inflow_outflow_rate_t3m	0.0209
		credit_lines_cnt	0.0208
		business_structure_partnership	0.0206
		loan_amount	0.0200
		fin_revenue	0.0198
		plat_saving_acct_flag	0.0195
		fin_liabilities	0.0190
		credit_inquiry_cnt_t12m	0.0188
		drv_tran_vol_std_t12m	0.0187
		drv_credit_inquiry_prop_t3t12	0.0180
		tran_outflow_t3m	0.0178
		plat_login_cnt_t1m	0.0162
		business_location_rural	0.0158

		drv_asset_liability_rate	0.0157
		drv_inflow_std_t6m	0.0154
		credit_current_amt	0.0151
		plat_dashboard_cnt_t3m	0.0137
		credit_inquiry_cnt_t3m	0.0131