Credit Data Analyst – Technical Task 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%):

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. 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,

c. 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.

d. 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, considering that default customers are only 8%, say 800 bad customers

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)

Hyperparameters Tuned:

- 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)

Hyperparameters Tuned:

- 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)

Hyperparameters Tuned:

- 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.

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.

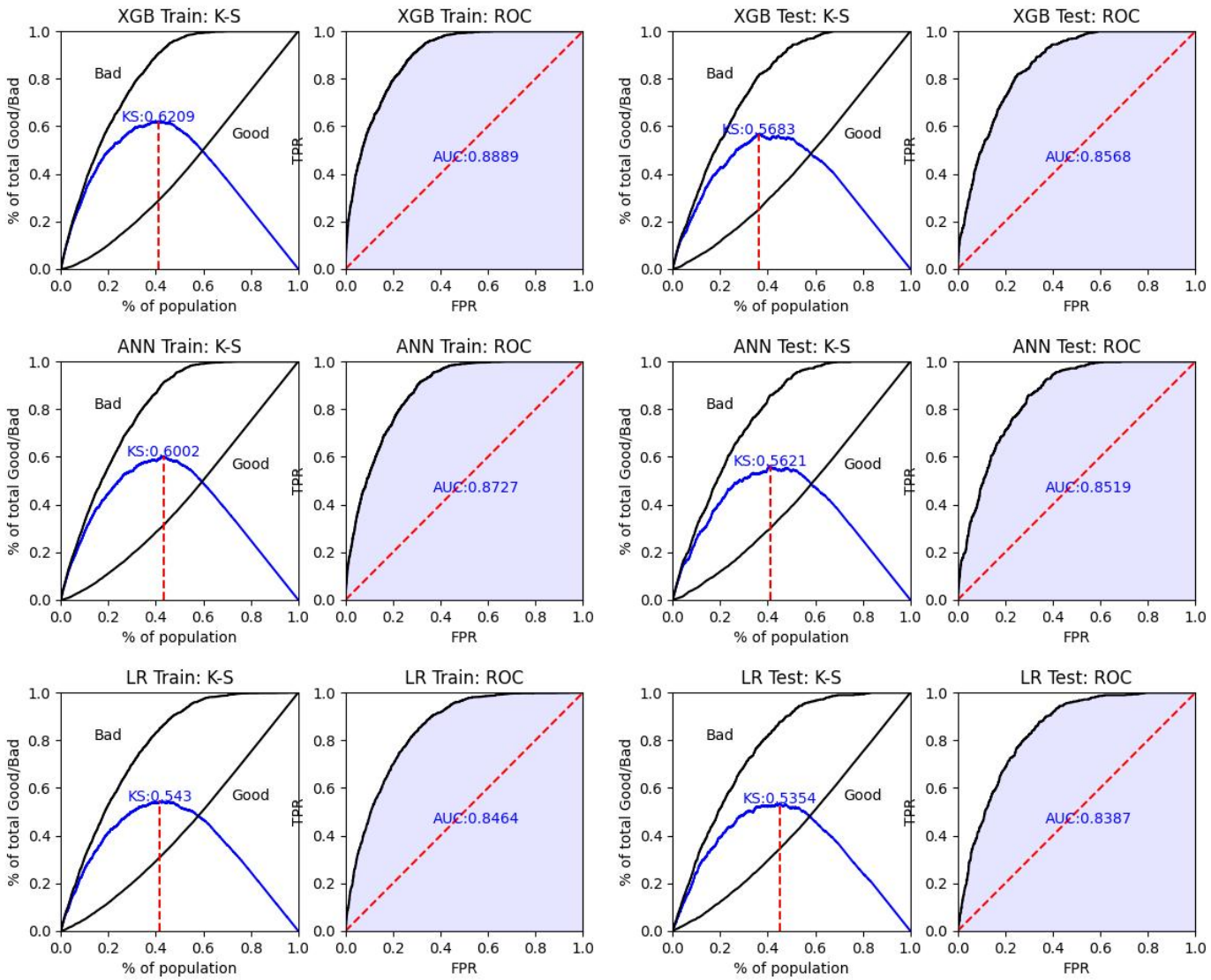
Rationale

- 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.

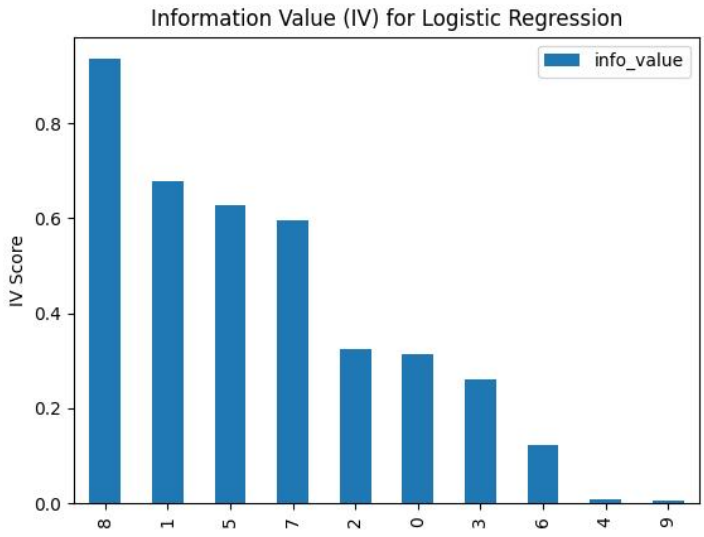
4.2. Model Performance

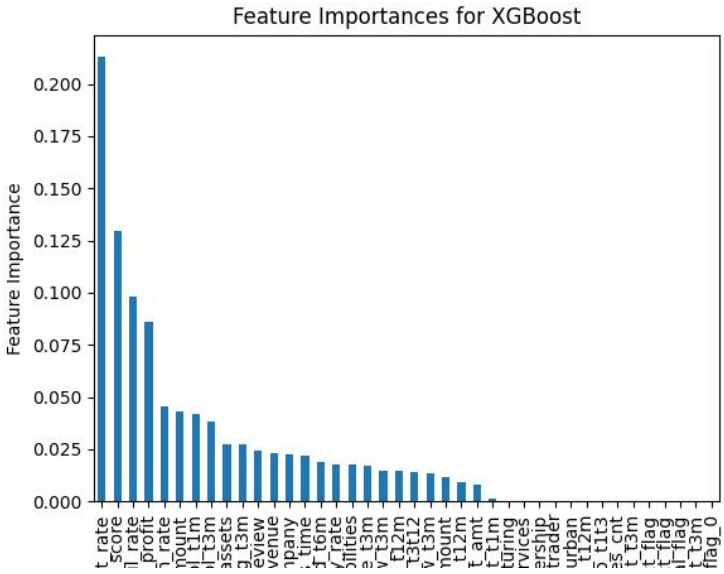
Model	Train AUC	Test AUC	Train KS	Test KS
LR	0.850	0.840	0.520	0.510
XGB	0.920	0.910	0.630	0.620
ANN	0.930	0.915	0.640	0.625

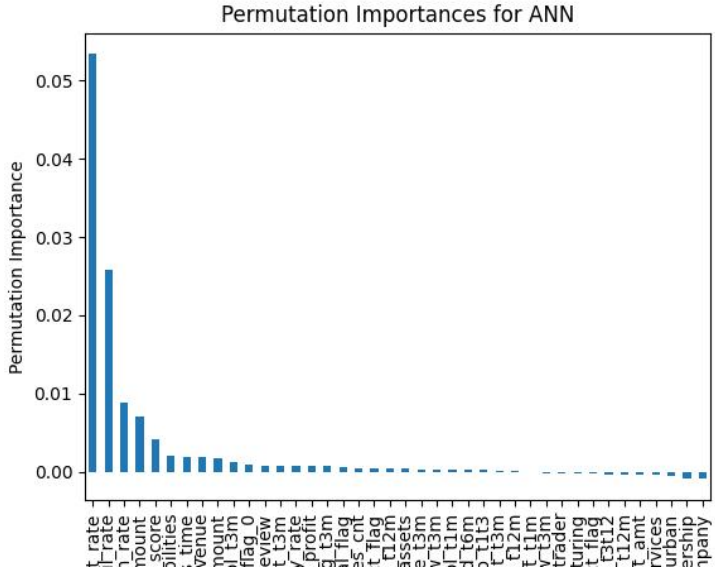
in Plotting:



4.3. Feature Importance







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. Key Drivers of Credit Risk (Soley based on auto-generated data, may have no real indications)
5.1.1.

5.2. Actionable Insights (based on key drives about)

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.	