This is a copy of a conversation between ChatGPT & Anonymous.

Report conversation

would this data dictionary content look better as a table? The analysis focused on features available at the time of the loan application to predict a binary loan default outcome (typically coded as 1 for default/charged off). Key predictive features identified through feature selection included: • EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3: External risk scores (i.e., credit scores), which were found to be consistent and strong predictors. • DAYS_BIRTH: Borrower age, which I converted into an interpretable variable. • AMT_CREDIT / loan_amnt: Measures of the loan size. • annual_inc / total income: Measures of borrower income. • INCOME_CREDIT_RATIO: A feature I engineered, calculated as total income divided by credit amount, designed to capture borrower leverage.

Yes. This content reads more clearly as a table because it standardizes naming, reduces repetition, and makes it easier to scan.

Feature (as used)	What it represents	Notes for your data dictionary
EXT_SOURCE_1	External risk score	One of three external risk score variables, strong predictor
EXT_SOURCE_2	External risk score	Strong predictor
EXT_SOURCE_3	External risk score	Strong predictor
DAYS_BIRTH	Borrower age	Stored as days, converted to an interpretable age variable
AMT_CREDIT / loan_amnt	Loan size	Different datasets use different field names, same concept
annual_inc / total income	Borrower income	Different datasets use different field names, same concept
INCOME_CREDIT_RATIO	Income-to-credit ratio	Engineered feature, calculated as total income ÷ credit amount, proxy for leverage

If you want it to look even cleaner, pick one naming convention per dataset and add a separate column like “Dataset field name” (or split into two tables if these are coming from different sources).

show example

Example table format for a data dictionary

Variable name	Data type	When available	Definition	How I used it
EXT_SOURCE_1	Numeric	At application	External risk score from a third-party source	Used as a core predictor, higher or lower values aligned with default risk
EXT_SOURCE_2	Numeric	At application	External risk score from a third-party source	Used as a core predictor, strong signal in feature selection
EXT_SOURCE_3	Numeric	At application	External risk score from a third-party source	Used as a core predictor, consistent with EXT_SOURCE_1 and EXT_SOURCE_2
DAYS_BIRTH	Integer	At application	Applicant age stored as days (typically negative), measured relative to application date	Converted to age in years for interpretability
AMT_CREDIT	Numeric	At application	Total credit amount requested or granted for the loan	Used as a loan size predictor
loan_amnt	Numeric	At application	Loan amount (alternate field name used in a different dataset)	Treated as the loan size measure when AMT_CREDIT was not present
annual_inc	Numeric	At application	Annual income (borrower-reported)	Used as an income predictor
total income	Numeric	At application	Total income (alternate field name used in a different dataset)	Treated as the income measure when annual_inc was not present
INCOME_CREDIT_RATIO	Numeric	Engineered at application	Total income divided by credit amount	Engineered feature used to capture borrower leverage

If you want, I can also show a version where AMT_CREDIT and loan_amnt are combined into one row (same for income), with an extra column called “Field name by dataset.”

Please format into a table I can insert into my project.

Variable name	Data type	When available	Definition	How I used it
EXT_SOURCE_1	Numeric	At application	External risk score from a third-party source	Included as a key predictor based on feature selection results
EXT_SOURCE_2	Numeric	At application	External risk score from a third-party source	Included as a key predictor based on feature selection results
EXT_SOURCE_3	Numeric	At application	External risk score from a third-party source	Included as a key predictor based on feature selection results
DAYS_BIRTH	Integer	At application	Borrower age recorded in days (often negative in the raw data)	Converted into an age-in-years variable to improve interpretability
AMT_CREDIT / loan_amnt	Numeric	At application	Loan amount or credit amount, depending on dataset naming	Used as the loan size measure
annual_inc / total income	Numeric	At application	Borrower income measure, depending on dataset naming	Used as the primary income measure
INCOME_CREDIT_RATIO	Numeric	Engineered at application	Total income divided by credit amount	Engineered feature used to capture borrower leverage relative to loan size

in markdown please

Variable name	Data type	When available	Definition	How I used it
EXT_SOURCE_1	Numeric	At application	External risk score from a third-party source	Included as a key predictor based on feature selection results
EXT_SOURCE_2	Numeric	At application	External risk score from a third-party source	Included as a key predictor based on feature selection results
EXT_SOURCE_3	Numeric	At application	External risk score from a third-party source	Included as a key predictor based on feature selection results
DAYS_BIRTH	Integer	At application	Borrower age recorded in days (often negative in the raw data)	Converted into an age-in-years variable to improve interpretability

Variable name	Data type	When available	Definition	How I used it
AMT_CREDIT / loan_amnt	Numeric	At application	Loan amount or credit amount, depending on dataset naming	Used as the loan size measure
annual_inc / total income	Numeric	At application	Borrower income measure, depending on dataset naming	Used as the primary income measure
INCOME_CREDIT_RATIO	Numeric	Engineered at application	Total income divided by credit amount	Engineered feature used to capture borrower leverage relative to loan size

markdown formatting that I can paste into a notebook

```
| Variable name | Data type | When available | Definition | How I used it |
|---|---|---|---|---|
| EXT_SOURCE_1 | Numeric | At application | External risk score from a third-party source | Included as a key predictor based on feature selection results |
| EXT_SOURCE_2 | Numeric | At application | External risk score from a third-party source | Included as a key predictor based on feature selection results |
| EXT_SOURCE_3 | Numeric | At application | External risk score from a third-party source | Included as a key predictor based on feature selection results |
| DAYS_BIRTH | Integer | At application | Borrower age recorded in days (often negative in the raw data) | Converted into an age-in-years variable to improve interpretability |
| AMT_CREDIT / loan_amnt | Numeric | At application | Loan amount or credit amount, depending on dataset naming | Used as the loan size measure |
| annual_inc / total income | Numeric | At application | Borrower income measure, depending on dataset naming | Used as the primary income measure |
| INCOME_CREDIT_RATIO | Numeric | Engineered at application | Total income divided by credit amount | Engineered feature used to capture borrower leverage relative to loan size |
```

Please include these if not already listed: Variable Name Data Type When Available Definition How I Used It Target Variable Binary Post-application (Outcome) Indicates whether a loan resulted in default (1) or non-default (0) (e.g., loan_default, TARGET) The variable I aimed to predict; used to stratify train-test splits. EXT_SOURCE_1, 2, 3 Numeric At application External risk scores (credit scores) from third-party sources. Identified as the strongest predictors of default risk. DAYS_BIRTH Integer At application Applicant age stored as days (typically negative). I converted this into an interpretable age variable for modeling. AMT_CREDIT / loan_amnt Numeric At application Total credit amount requested or granted for the loan. Used as a loan size measure. annual_inc / total income Numeric At application Borrower income measure. Used as a primary income predictor. INCOME_CREDIT_RATIO Numeric Engineered at application Calculated as total income divided by credit amount. I engineered this feature to capture borrower leverage.

```
| Variable name | Data type | When available | Definition | How I used it |
|---|---|---|---|---|
```

Target Variable (loan_default / TARGET)	Binary	Post-application (Outcome)	Indicates whether a loan resulted in default (1) or non-default (0).	The variable I aimed to predict; used to stratify train-test splits.
EXT_SOURCE_1	Numeric	At application	External risk score (credit score) from a third-party source.	Identified as one of the strongest predictors of default risk.
EXT_SOURCE_2	Numeric	At application	External risk score (credit score) from a third-party source.	Identified as one of the strongest predictors of default risk.
EXT_SOURCE_3	Numeric	At application	External risk score (credit score) from a third-party source.	Identified as one of the strongest predictors of default risk.
DAYS_BIRTH	Integer	At application	Applicant age stored as days (typically negative).	I converted this into an interpretable age variable for modeling.
AMT_CREDIT / loan_amnt	Numeric	At application	Total credit amount requested or granted for the loan.	Used as a loan size measure.
annual_inc / total income	Numeric	At application	Borrower income measure.	Used as a primary income predictor.
INCOME_CREDIT_RATIO	Numeric	Engineered at application	Calculated as total income divided by credit amount.	I engineered this feature to capture borrower leverage.

re-write the "How I used it" section so it is consistent. For example, some used "I..." while others do not.

Variable name	Data type	When available	Definition	How I used it
Target Variable (loan_default / TARGET)	Binary	Post-application (Outcome)	Indicates whether a loan resulted in default (1) or non-default (0).	I used this as the outcome variable to train and evaluate the model and to stratify train-test splits.
EXT_SOURCE_1	Numeric	At application	External risk score (credit score) from a third-party source.	I used this as a primary predictor of default risk based on feature selection results.
EXT_SOURCE_2	Numeric	At application	External risk score (credit score) from a third-party source.	I used this as a primary predictor of default risk based on feature selection results.
EXT_SOURCE_3	Numeric	At application	External risk score (credit score) from a third-party source.	I used this as a primary predictor of default risk based on feature selection results.
DAYS_BIRTH	Integer	At application	Applicant age stored as days (typically negative).	I converted this into an interpretable age variable and used it as a predictor in modeling.
AMT_CREDIT / loan_amnt	Numeric	At application	Total credit amount requested or granted for the loan.	I used this as a loan size predictor.
annual_inc / total income	Numeric	At application	Borrower income measure.	I used this as an income predictor.
INCOME_CREDIT_RATIO	Numeric	Engineered at application	Calculated as total income divided by credit amount.	I engineered this feature and used it to capture borrower leverage.

condense these phrases as they are going to be park of a table: I used this as a primary predictor of default risk based on feature selection results. |

I used this as a primary predictor of default risk.

update the entire section to be shorter explanations so they look good in a table

Variable name	Data type	When available	Definition	How I used it
Target Variable (loan_default / TARGET)	Binary	Post-application (Outcome)	Indicates whether a loan resulted in default (1) or non-default (0).	Outcome variable for prediction and evaluation.
EXT_SOURCE_1	Numeric	At application	External risk score (credit score) from a third-party source.	Primary predictor of default risk.
EXT_SOURCE_2	Numeric	At application	External risk score (credit score) from a third-party source.	Primary predictor of default risk.
EXT_SOURCE_3	Numeric	At application	External risk score (credit score) from a third-party source.	Primary predictor of default risk.
DAYS_BIRTH	Integer	At application	Applicant age stored as days (typically negative).	Converted to age in years and used as a predictor.
AMT_CREDIT / loan_amnt	Numeric	At application	Total credit amount requested or granted for the loan.	Loan size predictor.
annual_inc / total income	Numeric	At application	Borrower income measure.	Income predictor.
INCOME_CREDIT_RATIO	Numeric	Engineered at application	Calculated as total income divided by credit amount.	Engineered leverage predictor.

Please convert this to a table as well: Variable Name Description Role in Analysis TARGET / SeriousDlqin2yrs Binary flag (1 = Default, 0 = Repaid). The dependent variable (Target) for all supervised models. EXT_SOURCE_1 / 2 / 3 Normalized scores from external credit bureaus. Top predictors in Gradient Boosting; proxies for credit score. RevolvingUtilization Ratio of total balance to credit limit. Key indicator of financial stress in the KNN model. DAYS_BIRTH Age of the client in days (at application). Used to engineer the polynomial feature DAYS_BIRTH_SQ. AMT_INCOME_TOTAL Annual income of the applicant. Used to calculate Debt-to-Income (DTI) and leverage ratios. inq_last_6mths Number of credit inquiries in the last 6 months. Behavioral proxy for "credit seeking" behavior.

Variable name	Description	Role in analysis
TARGET / SeriousDlqin2yrs	Binary flag (1 = Default, 0 = Repaid).	Dependent variable for supervised models.
EXT_SOURCE_1 / 2 / 3	Normalized scores from external credit bureaus.	Top predictors in Gradient Boosting; credit risk proxies.
RevolvingUtilization	Ratio of total balance to credit limit.	Indicator of financial stress in the KNN model.
DAYS_BIRTH	Age of the client in days (at application).	Used to engineer polynomial feature DAYS_BIRTH_SQ.
AMT_INCOME_TOTAL	Annual income of the applicant.	Used to compute DTI and leverage ratios.
inq_last_6mths	Number of credit inquiries in the last 6 months.	Behavioral proxy for credit-seeking activity.

Variable Name,Description,Role in Analysis TARGET / SeriousDlqin2yrs,"Binary flag (1 = Default, 0 = Repaid).",The dependent variable (Target) for all supervised models. EXT_SOURCE_1 / 2 / 3,Normalized scores from external credit bureaus.,Top predictors in Gradient Boosting; proxies for credit score. RevolvingUtilization,Ratio of total balance to credit limit.,Key indicator of financial stress in the KNN model. AMT_INCOME_TOTAL,Annual income of the applicant.,Primary measure of repayment capacity. AMT_CREDIT,Final loan amount of the loan.,Used to determine total exposure and leverage. AMT_ANNUITY,Loan annuity (annual payment amount).,Critical for calculating debt service coverage. INCOME_CREDIT_RATIO,Engineered Feature: Total Income / Credit Amount.,"A proxy for ""Borrower Leverage."" Higher values indicate stronger repayment ability." AGE,Borrower age in years (derived from DAYS_BIRTH).,Used to analyze lifecycle risk; non-linear relationship with default. inq_last_6mths,Number of credit inquiries in the last 6 months.,"Behavioral proxy for ""credit seeking"" behavior."

Variable name	Description	Role in analysis
TARGET / SeriousDlqin2yrs	Binary flag (1 = Default, 0 = Repaid).	Dependent variable for supervised models.
EXT_SOURCE_1 / 2 / 3	Normalized scores from external credit bureaus.	Top predictors in Gradient Boosting; credit score proxies.
RevolvingUtilization	Ratio of total balance to credit limit.	Financial stress indicator in the KNN model.
AMT_INCOME_TOTAL	Annual income of the applicant.	Primary measure of repayment capacity.
AMT_CREDIT	Final loan amount.	Used to measure exposure and leverage.
AMT_ANNUITY	Loan annuity (annual payment amount).	Used to compute debt service burden.
INCOME_CREDIT_RATIO	Engineered feature: total income ÷ credit amount.	Borrower leverage proxy; higher implies stronger repayment capacity.
AGE	Borrower age in years (derived from DAYS_BIRTH).	Used to assess lifecycle risk and non-linear default patterns.
inq_last_6mths	Number of credit inquiries in the last 6 months.	Credit-seeking behavior proxy.

please convert into markdown: Model Type Specific Model(s) Dataset Rationale / Contribution Linear Regression Simple, Polynomial, Lasso, Ridge, Elastic Net Home Credit Default Risk Established a baseline and confirmed that linear relationships were too weak to capture the complexity of borrower behavior (R^2 ~0.006). Polynomial terms were used to test for non-linear relationships, but R^2 remained low. Classification (Foundational) Logistic Regression Home Credit Default Risk Showcased the challenge of class imbalance and the necessity of tuning, successfully adjusting model focus using class_weight='balanced' to capture 70% of actual defaults (Recall: 0.70). Classification (Distance-Based) Support Vector Machines (LinearSVC, RBF Kernel), K-Nearest Neighbors (KNN) Give Me Some Credit Explored models sensitive to distance and non-linearity. The RBF SVM showed that nonlinear decision boundaries were more effective than linear ones. KNN served as a baseline distance-based classifier. Classification (Ensemble/Tree-Based) Decision Tree, Random Forest, Gradient Boosting (GB) Lending Club, Home Credit Default Risk Tested non-linear methods. Random Forest achieved high accuracy but failed to identify defaults (nearly zero recall), highlighting limitations of accuracy in rare-event prediction. Gradient Boosting was chosen as the final superior predictive model. Unsupervised Clustering K-Means, DBSCAN, Hierarchical Agglomerative Clustering (HAC) Lending Club Used for borrower segmentation and revealing interpretable risk patterns independent of loan outcome. K-Means revealed two segments based on loan amount and interest rate. DBSCAN and HAC validated these clusters and identified outliers.

Model type	Specific model(s)	Dataset	Rationale / contribution
Linear regression	Simple, Polynomial, Lasso, Ridge, Elastic Net	Home Credit Default Risk	Established a baseline and showed linear effects were weak for borrower behavior ($R^2 \approx 0.006$). Polynomial terms tested non-linearity, but R^2 remained low.
Classification (foundational)	Logistic Regression	Home Credit Default Risk	Demonstrated class imbalance challenges and the need for tuning. Using `class_weight='balanced'` improved default capture (Recall = 0.70).
Classification (distance-based)	Support Vector Machines (LinearSVC, RBF), K-Nearest Neighbors (KNN)	Give Me Some Credit	Tested distance-sensitive,

non-linear models. RBF SVM outperformed linear boundaries. KNN provided a baseline distance-based classifier. |
| Classification (ensemble/tree-based) | Decision Tree, Random Forest, Gradient Boosting (GB) | Lending Club; Home Credit Default Risk | Evaluated non-linear methods. Random Forest had high accuracy but near-zero default recall, highlighting rare-event limitations. Gradient Boosting performed best and was selected as the final model. |
| Unsupervised clustering | K-Means, DBSCAN, Hierarchical Agglomerative Clustering (HAC) | Lending Club | Segmented borrowers without using outcomes. K-Means produced two segments driven by loan amount and interest rate. DBSCAN and HAC supported the segmentation and flagged outliers. |

Model Type,Specific Model(s),Dataset,Rationale & Contribution Linear Regression,"Simple, Polynomial, Lasso, Ridge, Elastic Net",Home Credit,"Established a baseline and confirmed that linear relationships were too weak to capture borrower behavior complexity. Even with polynomial terms, variance explanation remained negligible, necessitating non-linear approaches." Classification (Foundational),Logistic Regression,Home Credit,"Highlighted the challenge of class imbalance. Standard logistic regression failed to detect defaults, but adjusting the class_weight='balanced' parameter proved that the model could be forced to focus on the minority class." Classification (Distance-Based),"Support Vector Machines (RBF), K-Nearest Neighbors (KNN)",Give Me Some Credit,"Explored models sensitive to geometric distance. The RBF Kernel SVM demonstrated that non-linear decision boundaries were superior to linear ones, while KNN served as a non-parametric baseline." Classification (Ensemble),"Decision Tree, Random Forest, Gradient Boosting","Lending Club, Home Credit","Tested robust non-linear methods. While Random Forest struggled with rare events, Gradient Boosting was selected as the final champion model due to its ability to sequentially correct errors on difficult examples." Unsupervised Clustering,"K-Means, DBSCAN, Hierarchical (HAC)",Lending Club,"Used for borrower segmentation to reveal interpretable risk patterns independent of loan outcome. K-Means identified leverage-based segments, while DBSCAN was utilized specifically for outlier detection."

Model type	Specific model(s)	Dataset	Rationale & contribution
Linear Regression	Simple, Polynomial, Lasso, Ridge, Elastic Net	Home Credit	Established a baseline and showed linear relationships were too weak to capture borrower behavior complexity. Polynomial terms did not improve explanatory power, supporting a shift to non-linear methods.
Classification (Foundational)	Logistic Regression	Home Credit	Highlighted class imbalance. Baseline logistic regression under-detected defaults, but `class_weight='balanced'` shifted focus to the minority class.
Classification (Distance-Based)	Support Vector Machines (RBF), K-Nearest Neighbors (KNN)	Give Me Some Credit	Tested distance-based models. RBF SVM showed non-linear decision boundaries outperformed linear ones. KNN served as a non-parametric baseline.
Classification (Ensemble)	Decision Tree, Random Forest, Gradient Boosting	Lending Club; Home Credit	Evaluated non-linear ensembles. Random Forest struggled with rare events, while Gradient Boosting was selected as the final model due to sequential error-correction on difficult cases.
Unsupervised Clustering	K-Means, DBSCAN, Hierarchical (HAC)	Lending Club	Segmented borrowers without outcome labels. K-Means identified leverage-based segments. DBSCAN supported outlier detection. HAC provided cluster validation.

please format it so it looks more like a table taht is easy to read.

| Model type | Specific model(s) | Dataset | Rationale and contribution |

```
|---|---|---|---|
| Linear regression | Simple, Polynomial, Lasso, Ridge, Elastic Net | Home
Credit | Baseline comparison. Linear effects were weak and polynomial terms
did not improve fit, so non-linear methods were needed. |
| Classification (foundational) | Logistic Regression | Home Credit | Class
imbalance test. Baseline detection of defaults was weak, but
`class_weight='balanced'` shifted focus to the minority class. |
| Classification (distance-based) | SVM (RBF), KNN | Give Me Some Credit |
Distance-based modeling. RBF SVM showed non-linear boundaries outperformed
linear ones. KNN provided a non-parametric baseline. |
| Classification (ensemble) | Decision Tree, Random Forest, Gradient Boosting
| Lending Club; Home Credit | Robust non-linear methods. Random Forest
struggled with rare events. Gradient Boosting performed best and was selected
as the final model. |
| Unsupervised clustering | K-Means, DBSCAN, Hierarchical (HAC) | Lending
Club | Borrower segmentation. K-Means identified leverage-based segments.
DBSCAN flagged outliers. HAC supported cluster validation. |
```
for apa formatting, do I label "table" before or after the actual table?

In APA 7, label it before the table.

Put the table number on one line, then the title on the next line, then the table.

Example:

Table 1
Modeling Approaches, Datasets, and Contributions

[table here]

If you are using a Jupyter notebook, you can place the table label and title in the markdown cell directly above the markdown table.

justified left or center?

APA 7 uses left alignment for the table number and title.

- Table number: left, bold (e.g., **Table 1**)
- Title: left, italic (title case), on the next line
- Table itself: left-aligned within the page margins

In a Jupyter notebook markdown cell, you can keep the label and title left-aligned.

same question, but for figuires

For APA 7 figures, use the same approach.

Place the figure number and title before the figure, left-aligned.

- Figure number: left, bold (for example, **Figure 1**)
- Title: left, italic (title case), on the next line
- Figure image: below the title
- Figure note, if needed: below the figure (left-aligned), starting with “Note.”

Example:

Figure 1
ROC Curve for Gradient Boosting Model

[figure here]

Note. Brief explanation of the figure, abbreviations, or data details.

can you export this picture into text?
Thought for 1m 12s

Model	Accuracy	Recall (Defaulters)	F1-Score (Defaulters)	Key Observation
Random Forest (Balanced)	0.855	0.00	0.01	Failed, predicted "No Default" for nearly all cases despite class weighting.
K-Nearest Neighbors (k=21)	0.936	0.09	0.16	Conservative, high precision but missed >90% of defaults.
Gradient Boosting (Default)	~92.0%	0.01	0.02	Baseline, poor sensitivity at the standard 0.50 threshold.
Gradient Boosting (Tuned)	N/A*	0.40	0.29	Champion, threshold tuning significantly improved default detection.

apa cite: https://www.cotality.com/press-releases/late-stage-missed-payments-raise-delinquency-rates
Thought for 12s

Cotality. (2025, December 9). Late-stage missed payments raise delinquency rates [Press release]. https://www.cotality.com/press-releases/late-stage-missed-payments-raise-delinquency-rates

cotality.com

In-text citation: (Cotality, 2025) cotality.com