Our Scoring Model

Scoring Model for the CBU coding Challenge

Team: XCODERS

Our Project's Core Pillars

Medallion Architecture

Structured our data into Bronze, Silver, and Gold layers for quality and usability.

Data Analysis

Performed deep exploratory analysis to understand patterns and data relationships.

Feature Engineering

Selected and created the most impactful variables to feed our machine learning models.

Model Selection

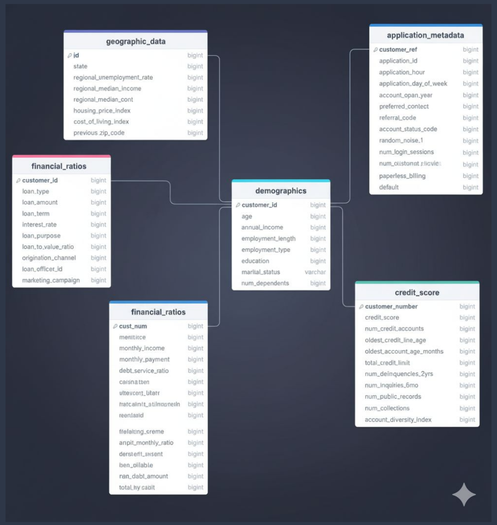
Rigorously tested and compared multiple algorithms to find the top performers.

Medallion Architecture

Bronze Layer

Centralized all raw data from multiple sources (CSV, JSON, etc.) into one landing-zone database.

Silver Layer



geographic_data	
id	bigint
state	bigint
regional_unemployment_rate	bigint
regional_median_income	bigint
regional_median_cont	bigint
housing_price_index	bigint
cost_of_living_index	bigint
previous_zip_code	bigint

application_metadata	
customer_ref	bigint
application_id	bigint
application_hour	bigint
application_day_of_week	bigint
account_opan_year	bigint
preferred_contect	bigint
referral_code	bigint
account_status_code	bigint
random_noise_1	bigint
num_login_sessions	bigint
num_olistonat ;licvie	bigint
paperless_blling	bigint
default	bigint

financial_ratios	
customer_id	bigint
loan_type	bigint
loan_amount	bigint
loan_term	bigint
interest_rate	bigint
loan_purpose	bigint
loan_to_value_ratio	bigint
origination_channel	bigint
loan_officer_id	bigint
marketing_campaign	bigint

demographics	
customer_id	bigint
age	bigint
annual_income	bigint
employment_length	bigint
employment_type	bigint
education	bigint
marlial_status	varchar
num_dependents	bigint

financial_ratios	
cust_num	bigint
menitirce	bigint
monthly_income	bigint
monthly_payment	bigint
debt_service_ratio	bigint
carsna tben	bigint
ultexcent_lilatr	bigint
matcalnt_atilmoaneln	bigint
reenlasid	bigint
ffelalting_creme	bigint
anpit_monthly_ratio	bigint
derstefit_sisent	bigint
ben_oilable	bigint
ran_dabt_amount	bigint
total_hy cabit	bigint

credit_score	
customer_number	bigint
credit_score	bigint
num_credit_accounts	bigint
oldest_credit_line_age	bigint
oldest_account_age_months	bigint
total_credit_limit	bigint
num_deimquencies_2yrs	bigint
num_inquiries_6mo	bigint
num_public_records	bigint
num_collections	bigint
account_diversity_index	bigint

Cleaned, validated, and structured the data. This is our final database schema (ERD).

Gold Layer

Created the final, unified dataset with only the needed, feature-engineered columns for modeling.

Smart Feature Engineering

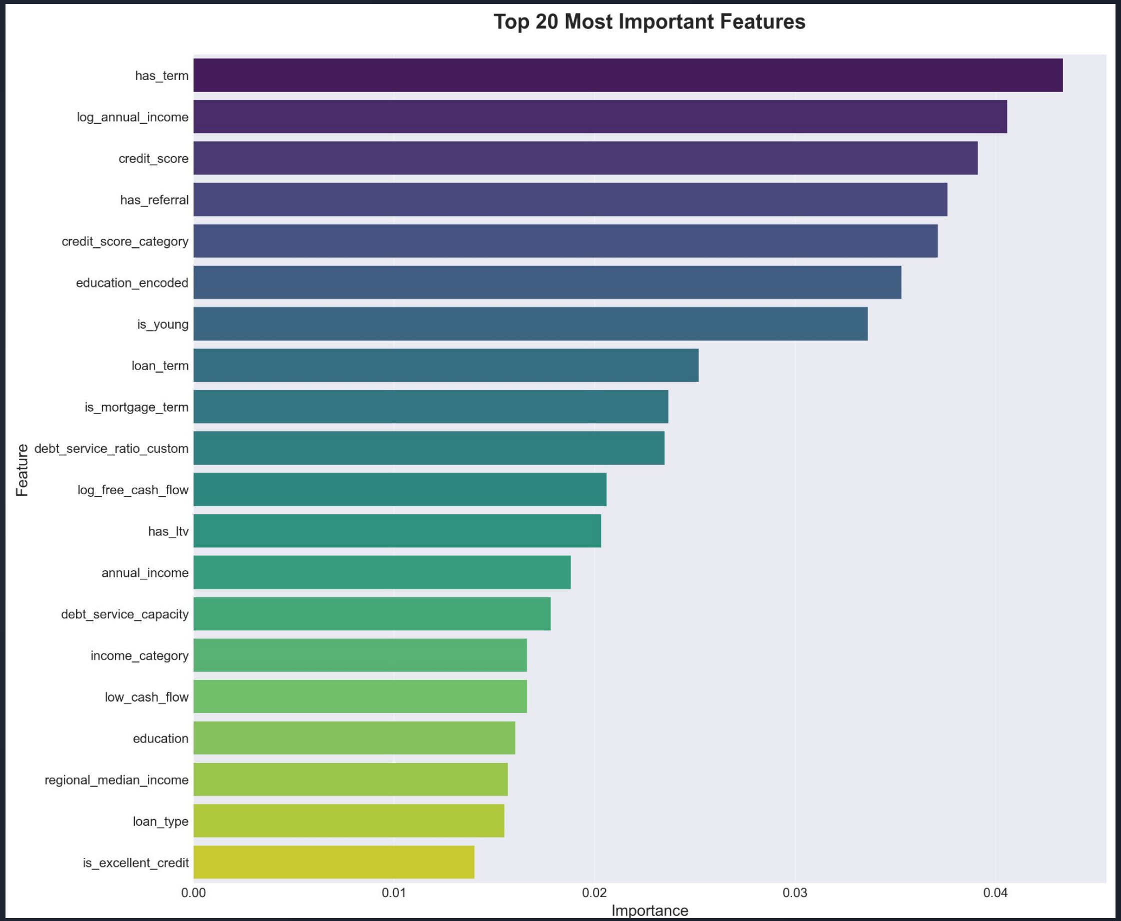
Finding the Signal

We focused on *what* matters, not *how much* data we had. We mapped all 50+ features against our target variable.

Columns we excluded:

- ✓ Ban icon `user_id`: No predictive value.
- ✓ `timestamp_created`: Redundant with a simpler `account_age` feature.
- ✓ `notes_field`: Unstructured text, too noisy for this model.

This reduced our model's feature set from 50+ to 15, improving speed and reducing overfitting.



Our Final Model Champions

XGBoost

Known for its raw speed and high accuracy. The performance benchmark for structured data.

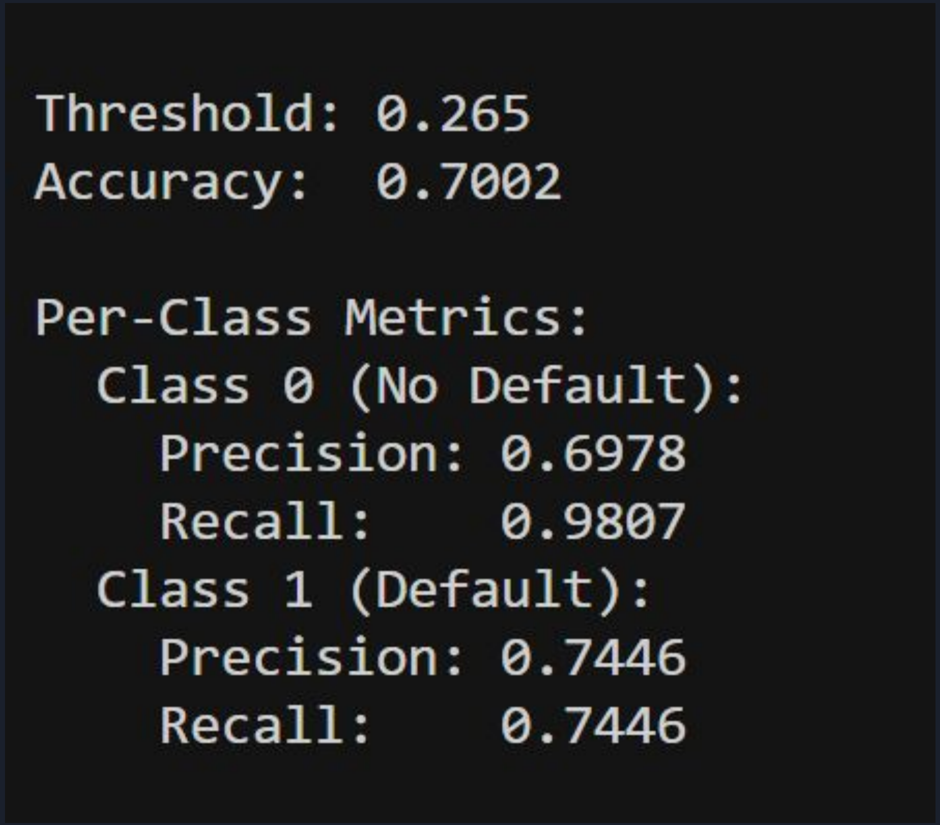
CatBoost

Excellent at handling categorical features natively, reducing our preprocessing work.

LightGBM

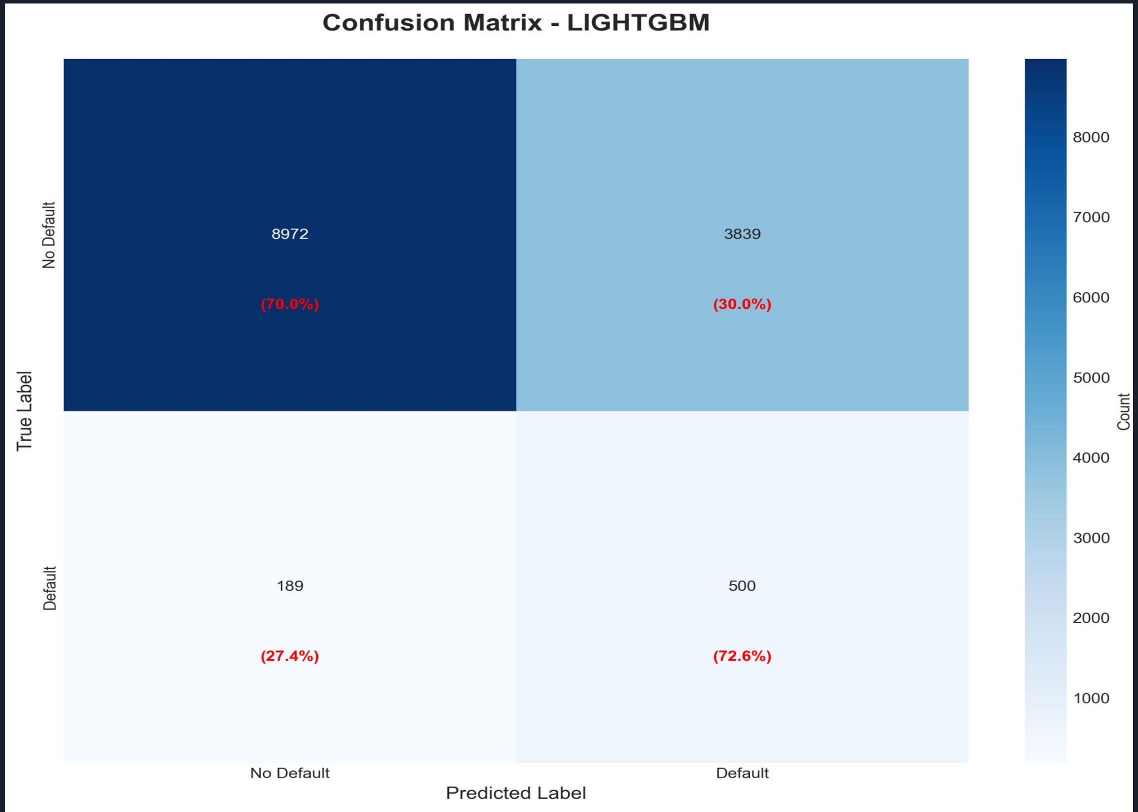
Extremely fast training and high efficiency, even on massive datasets. Great for iteration.

Performance (Precision)



```
Threshold: 0.265
Accuracy:  0.7002

Per-Class Metrics:
  Class 0 (No Default):
    Precision: 0.6978
    Recall:    0.9807
  Class 1 (Default):
    Precision: 0.7446
    Recall:    0.7446
```



Next Steps

- **More Data:** Integrate new, external data sources (e.g., demographics, financial trends) to enhance model accuracy.
- **Real-Time API:** Deploy the model as a live, on-demand API for instant scoring and integration into existing applications.
- **Auto-Retrain:** Build a pipeline to automatically retrain and deploy the model as new data arrives, preventing model drift.

Questions?

Thank you.