LendingClub

UPDATE - 2

JASMIN'S INVESTMENT PLAYBOOK: SMARTER BETS WITH PREDICTIVE POWER

Group 21:
Angela Pan
Poorvi Phulwani
Sai Priya Sure

FROM CURIOSITY TO STRATEGY

Last Week's Recap

Explored variables, risk-return hypotheses, cluster patterns

This Week's Focus

Predict loan defaults (classification)
Predict returns (regression)
Implement and compare investment strategies

POST-LOAN INFORMATION INFLATES MODEL PERFORMANCE, CREATING A MISLEADING IMPRESSION OF PREDICTIVE POWER

Some information in the dataset gives away the ending — like knowing if someone paid you back after the loan is over. While these make models look perfect, they're not useful for real-time decision-making.

What we did?

Built two machine learning models to predict if a loan will default:

- One with leakage features (like "how much money was recovered")
- One without leakage features (only info available at the time of loan application)

Model Type	Accuracy	AUC Score	Notes
With Leakage	100%	0.999	Too good to be true - knows the future!
Without Leakage	82%	0.666	Realistic performance using only pre-loan features

Variables That Cause Leakage:

Exclude these fields from predictive modeling as they're only available after loan closure.

- Total Payment Received (total_pymnt)
- Amount Recovered (recoveries)
- Principal Repaid (total_rec_prncp)
- Interest Repaid (total_rec_int)
- Late Fees Paid (total_rec_late_fee)

```
Scenario 1: Model WITH Leakage Features

--- With Leakage Model Performance ---
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     38696
           1       1.00      0.98      0.99      8365

    accuracy                           1.00     47061
   macro avg       1.00      0.99      0.99     47061
weighted avg       1.00      1.00      1.00     47061


ROC AUC Score: 0.9997110154454214

Scenario 2: Model WITHOUT Leakage Features

--- Without Leakage Model Performance ---
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.99      0.90     38696
           1       0.43      0.03      0.06      8365

    accuracy                           0.82     47061
   macro avg       0.63      0.51      0.48     47061
weighted avg       0.76      0.82      0.75     47061


ROC AUC Score: 0.666088636285279
```

For meaningful predictions, stick to information available at loan application time only — like income, loan amount, and debt-to-income ratio and exclude post-loan features to prevent signal leakage in default prediction models to build trustworthy and real-time loan default models.

LENDINGCLUB'S RISK ASSESSMENT VARIABLES (GRADE AND INTEREST RATE) PROVIDE USEFUL BUT INSUFFICIENT PREDICTIVE POWER FOR LOAN DEFAULT PREDICTION.

LendingClub assigns each loan a Grade (A–G) and Interest Rate based on their internal risk model. But are these two signals alone enough to predict default? **Short answer: Not really.**

What we did?

Trained two models using only:

- Loan Grade (LendingClub's risk label, A to G)
- Interest Rate (LendingClub's pricing decision)

Models Used:

- Logistic Regression
- Decision Tree Classifier

Model	Accuracy	ROC AUC Score	Observation
Logistic Regression	82%	0.676	Poor detection of defaults (almost 0%)
Decision Tree	82%	0.679	Similar results — default class ignored

Why This Matters (and What the Model Revealed):

- LendingClub's signals like Grade and Interest Rate do capture some loan risk, but they miss deeper borrower-level details like income, debt-to-income ratio, and credit history.
- In fact, the model relied almost entirely on Interest Rate (96% importance), with Grades B–G contributing less than 2.5% combined.

```
--- Analyzing Signals Generated by LendingClub ---

Logistic Regression with LendingClub Signals:

Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.82      1.00      0.90     38745
           1       0.00      0.00      0.00      8316

    accuracy                           0.82     47061
   macro avg       0.41      0.50      0.45     47061
weighted avg       0.68      0.82      0.74     47061


Logistic Regression ROC AUC Score: 0.6765770006413961

Decision Tree with LendingClub Signals:

Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.82      1.00      0.90     38745
           1       0.00      0.00      0.00      8316

    accuracy                           0.82     47061
   macro avg       0.41      0.50      0.45     47061
weighted avg       0.68      0.82      0.74     47061


Decision Tree ROC AUC Score: 0.6786564959490498

Decision Tree Feature Importances:
int_rate: 0.9628022443816818    grade_E: 7.700241480188192e-05
grade_B: 0.02433295795846398    grade_F: 0.0
grade_C: 0.002317508300188214   grade_G: 0.0045043089761547745
grade_D: 0.00596597796870911
```

LendingClub's signals are a helpful starting point, but they're not sufficient to reliably predict defaults—Jasmin needs deeper analysis using borrower-specific and pre-loan features.

DOWNSAMPLING THE NON-DEFAULTED LOANS SIGNIFICANTLY IMPROVES THE MODEL'S ABILITY TO IDENTIFY POTENTIAL DEFAULTS, THOUGH AT THE COST OF OVERALL ACCURACY.

In our dataset, only 18% of loans defaulted — meaning models trained as-is tend to ignore defaulted cases. To fix this, we used downsampling to balance the data and give defaults a fair voice.

What we did?

The dataset had:

- Non-defaults - 82%
- Defaults - 18%

We tested 3 sampling strategies with 3 models:

1. No Sampling 2. Auto Downsampling 3. Majority Class Downsampling

We tested three downsampling strategies:

- No Sampling: Train on full dataset as-is
- Auto Sampling: Let the model choose internal balance
- Majority Sampling: Force a 1:1 ratio by downsampling safe loans

Model	Strategy	Accuracy	ROC AUC
Logistic Regression	No Sampling	0.8236	0.6918
Decision Tree	No Sampling	0.7178	0.5322
Random Forest	No Sampling	0.8225	0.6746
Logistic Regression	Auto Sampling	0.6206	0.6913
Decision Tree	Auto Sampling	0.5563	0.5549
Random Forest	Auto Sampling	0.6285	0.6786
Logistic Regression	Majority Sampling	0.6206	0.6913
Decision Tree	Majority Sampling	0.5563	0.5549
Random Forest	Majority Sampling	0.6285	0.6786

What Changed After Downsampling?

- Downsampling significantly improved default detection, raising recall from nearly 0% to around 65%, making models useful for identifying risky loans.
- Models without downsampling were biased, favoring non-defaulted loans and achieving high accuracy at the cost of missing critical defaults.
- Logistic Regression consistently handled imbalance better, while Decision Trees struggled more without sampling.
- There's no one-size-fits-all solution — the effectiveness of downsampling depends on the model used, so strategies must be tailored accordingly.

Downsampling corrected model bias, boosting default detection and enabling more balanced, actionable predictions which are crucial for Jasmin's investment strategy despite lower overall accuracy.

FEATURE SELECTION: UNDERSTANDING PREDICTOR RELATIONSHIPS

We tested how well we could predict loan default using only pre-loan application features-no LendingClub-derived variables like Grade or Interest Rate. We checked how variables relate to each other — to avoid confusing our models with redundant signals (especially important for logistic regression).

What Features We Used:

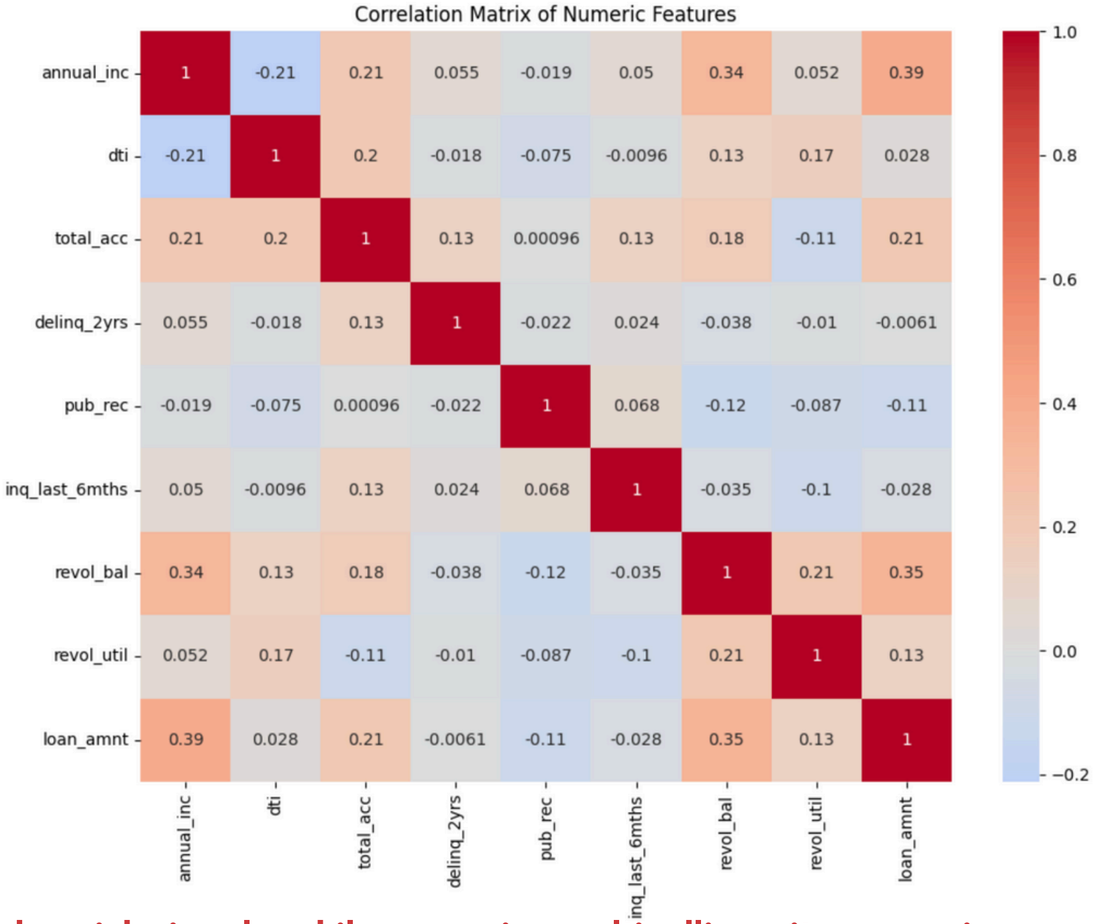
- Borrower Financials: Income, debt-to-income ratio, number of credit lines
- Credit History: Public records, inquiries, earliest credit date
- Loan Characteristics: Amount, term
- Categorical Info: Employment length, home ownership, purpose of loan, verification status

Correlation Highlights(circle on the right):

- loan amount & annual income → 0.39
- loan amount & revolving credit balance → 0.35
- revolving credit balance & annual income → 0.34

Key Insights:

- High correlation (above 0.7) can hurt model performance
- Here: All correlations are moderate → kept all features
- For Logistic Regression: applied scaling + L1 regularization



Decision Tree:	Logistic Regression:
0.53 AUC	0.66 AUC

Focus on term length, loan purpose, housing status, and credit history as key risk signals, while managing multicollinearity among income, revolving credit balance, and loan amount in regression models.

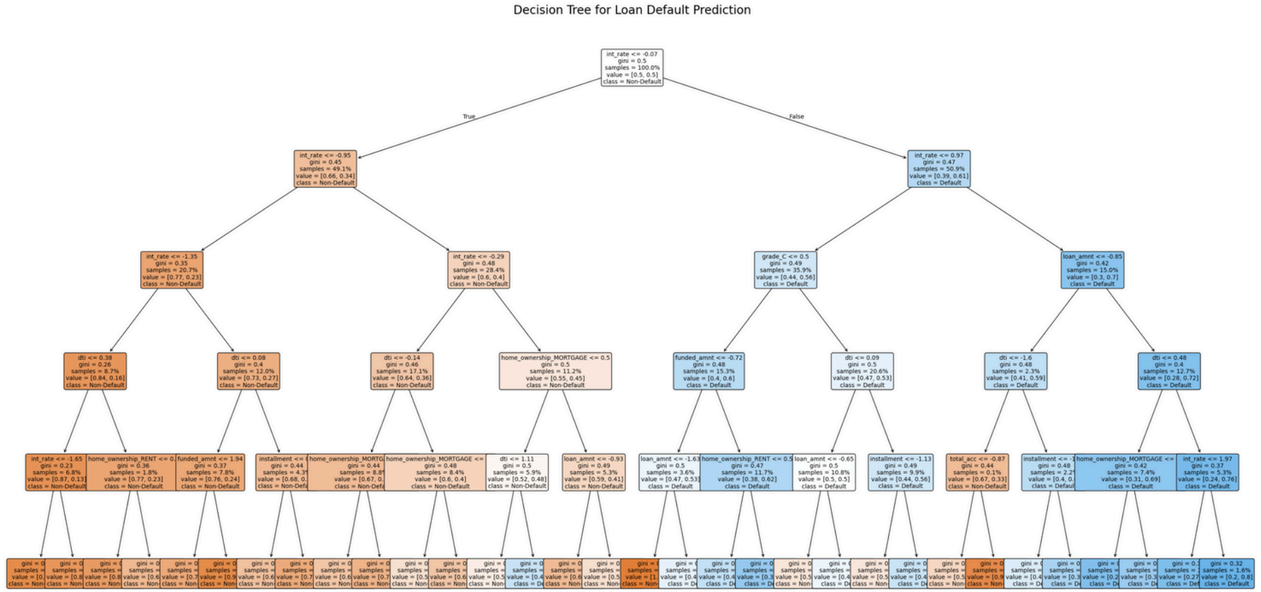
DECISION TREE ANALYSIS: CLEAR PATTERNS FOR INVESTMENT DECISIONS

We trained a model that works like a series of yes/no questions to decide whether a loan is risky. It learns patterns from past loans and uses them to guide new decisions.

Key Findings

Our decision tree model reveals distinct patterns that can guide Jasmin's investment strategy:

- Interest rate dominates default prediction (88% of model importance)
- Loan grade influences model splits, with Grade C being a key differentiator (1.7% importance)
- Income, debt-to-income ratio, and loan amount collectively contribute meaningful signals
  - Low DTI ratios correlate with safer loans
- Our model identifies ~70% of defaults (3,645 out of 5,197)
  - 89% precision for safe loans helps avoid unnecessary rejections




```
--- Decision Tree Performance ---

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.57      0.70     23451
           1       0.27      0.70      0.39      5197

    accuracy                           0.60     28648
   macro avg       0.58      0.64      0.54     28648
weighted avg       0.78      0.60      0.64     28648

Accuracy:  0.5958
Precision: 0.2666
Recall:    0.7014
F1 Score:  0.3863
ROC AUC:   0.6873

Confusion Matrix:
[[13423 10028]
 [ 1552  3645]]

Top Decision Rules:
If int_rate <= -0.9493
If int_rate <= -1.3483
If dti <= 0.3821

Top 10 Most Important Features:
int_rate: 0.8805
dti: 0.0338
loan_amnt: 0.0237
grade_C: 0.0171
home_ownership_MORTGAGE: 0.0139
funded_amnt: 0.0118
home_ownership_RENT: 0.0109
installment: 0.0075
total_acc: 0.0008
annual_inc: 0.0000
```

Use the decision tree to prioritize low-interest, Grade A loans with low DTI and well-sized amounts—it's a clear, rule-based tool that improves default avoidance over random or basic screening.

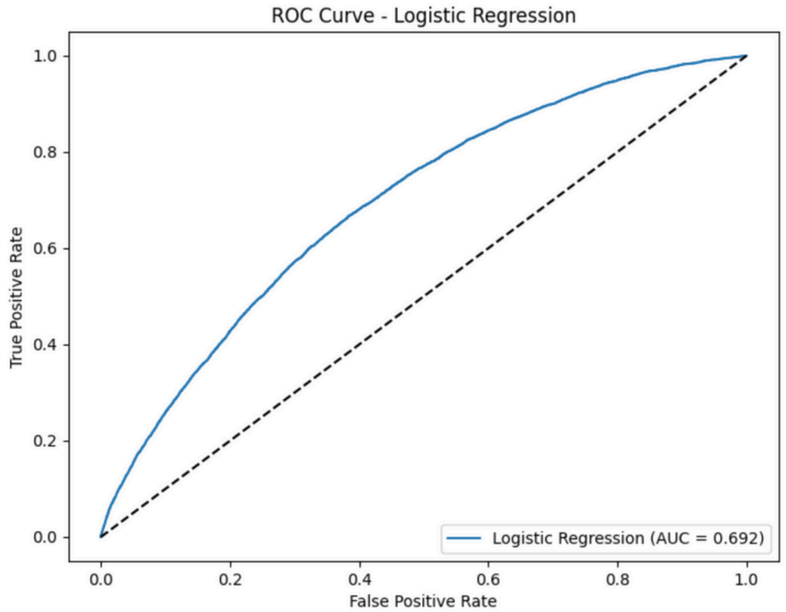
LOGISTIC REGRESSION: UNCOVERING DEFAULT RISK PATTERNS

What we did?

We built a model that looks at many loan details to estimate how likely someone is to default. Unlike a decision tree, this model doesn't follow rules — it calculates a risk score using all the numbers together.

Top Signal for Default Risk

Feature	Interpretation
Grade A loans	Top grade significantly reduces default risk
Interest rate	Higher interest → ↑ default risk
Funded amount	Larger loans increase default risk
Employment length missing	No job history info → ↑ risk
Installment	Lower monthly payments relative to loan size decrease risk
Grade B loans	Grade B also offers protection
home_ownership_MORTGAGE	Safer if borrower has a mortgage
home_ownership_RENT	Higher risk for renters
Annual income	Higher income reduces default probability



Model Performance

- Identifies 68% of defaults with 90% precision on safe loans
- 0.69 AUC shows good discriminative ability

Example Prediction

- Grade B loan at 9.17% interest rate → 28% default probability (predicted safe)
- Grade F loan at 24.5% interest rate → 71% default probability (predicted default)

This model sees through the numbers — like a credit whisperer. It weighs everything from interest rate to job history. It's not perfect, but it gives Jasmin a fair and explainable risk estimate.

MODEL COMPARISON: CHOOSING THE BEST APPROACH FOR LOAN INVESTMENT

Performance Comparison

Looking at both models side by side reveals similar predictive capabilities:

Metric	Decision Tree	Logistic Regression
Accuracy	61.5%	62.0%
Recall (Default)	66.1%	68.0%
Precision	26.4%	27.0%
ROC AUC	0.680	0.692

Both models demonstrate moderate discriminative power with **logistic regression** showing a slight edge.

Did Our Early Hypotheses Hold Up?

We guessed interest rate, loan amount, and income would matter most.
Turns out — we were mostly right!

- Interest rate is the dominant predictor for default risk in both the models
- Grades and debt ratio (DTI) also turned out to be strong signals
- Annual income had only minor effect

Key Differences between both Models

- Logistic regression provides probabilistic outputs suitable for ranking
- Decision tree offers simple, interpretable rules for quick screening
- Logistic regression captures more subtle relationships across features

Logistic Regression is our choice!

- Lets understand this using cost benefit analysis
  - Expected Return = P(default) × E(Return|default) + (1-P(default)) × E(Return|no default)

MODEL COMPARISON: CHOOSING THE BEST APPROACH FOR LOAN INVESTMENT

Assumptions

- For $1,000 invested across 100 loans with conservative assumptions:
- Non-defaulted loans return 7% ($10 → $10.70)
- Defaulted loans lose 50% ($10 → $5.00)

Decision Tree

Confusion Matrix →

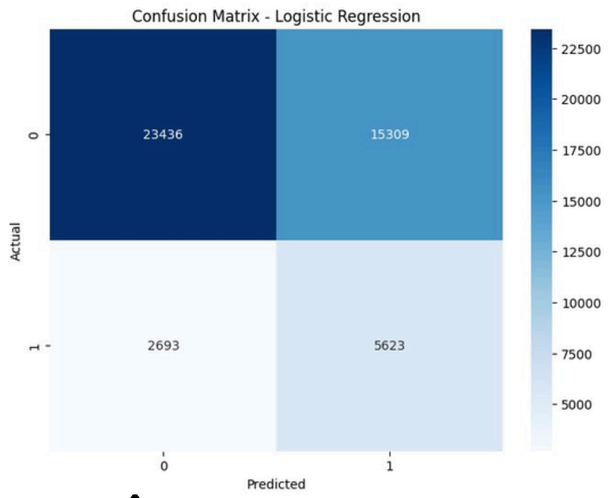
```
Confusion Matrix:
[[23429 15316]
 [ 2821  5495]]
```

Loans predicted as non-default (what we would invest in): 23,429 + 2,821 = 26,250
Of those, 23,429 were truly good loans = success rate = 23,429 / 26,250 ≈ 0.8926
Default rate on selected loans = 1 - 0.8926 = 0.1074

Expected Return Calculation

For each $10 investment:
Non-default return: $10 × 1.07 = $10.70 ; Default return: $10 × 0.50 = $5.00
Expected return per $10: = 0.8926 × $10.70 + 0.1074 × $5.00 = $9.55 + $0.54 = $10.09
Profit per $10: $0.09

For $1,000 invested across 100 loans: 100 × $0.09 = $9.00 profit



Logistic Regression

Confusion Matrix →

Loans predicted as non-default (what we would invest in): 23,436 + 2,693 = 26,129
Of those, 23,436 were truly good loans = success rate = 23,436 / 26,129 ≈ 0.8970
Default rate on selected loans = 1 - 0.8970 = 0.1030

Expected Return Calculation

For each $10 investment:
Non-default return: $10 × 1.07 = $10.70
Default return: $10 × 0.50 = $5.00
Expected return per $10: = 0.8970 × $10.70 + 0.1030 × $5.00 = $9.60 + $0.52 = $10.12
Profit per $10: $0.12

For $1,000 invested across 100 loans: 100 × $0.12 = $12.00 profit

RETURN PREDICTION: SIMPLER IS SOMETIMES BETTER

Which Return Formula Should We Use?

From Update 1, we explored 3 return measures, we choose intermediate return measure to use at this time for following reason:

- **Realistic risk assessment** - Neither overly optimistic (assuming full payment until default) nor pessimistic (assuming immediate total loss)

Return Measure	What It Represents
ret_a	Profit ÷ Investment
ret_b	Time-adjusted return
ret_c	Complex adjusted return

We trained return prediction models (e.g., Lasso, Ridge) for each return method and evaluated:

- RMSE (lower = better prediction accuracy)
- R^2 (higher = better explanation of variation)

And What We Found

- **ret_a had lowest RMSE and highest R^2 with Ridge**
- ret_b was okay but slightly less accurate
- ret_c underperformed on both metrics

We tested three ways to measure loan profit. The simplest one — just comparing earnings to the money invested — turned out to be not only the easiest to understand, but also the most accurate to predict! Based on these results, we recommend Jasmin use this straightforward return measurement (ret_a) for her prediction models

RANDOM BEATS MODELS: THE SURPRISING REALITY OF P2P LOAN SELECTION

Objective:

Use borrower and loan details available at application time to predict loan returns (%), aiding smarter investment decisions.

Models & Return Measures Tested:

Return Measures:

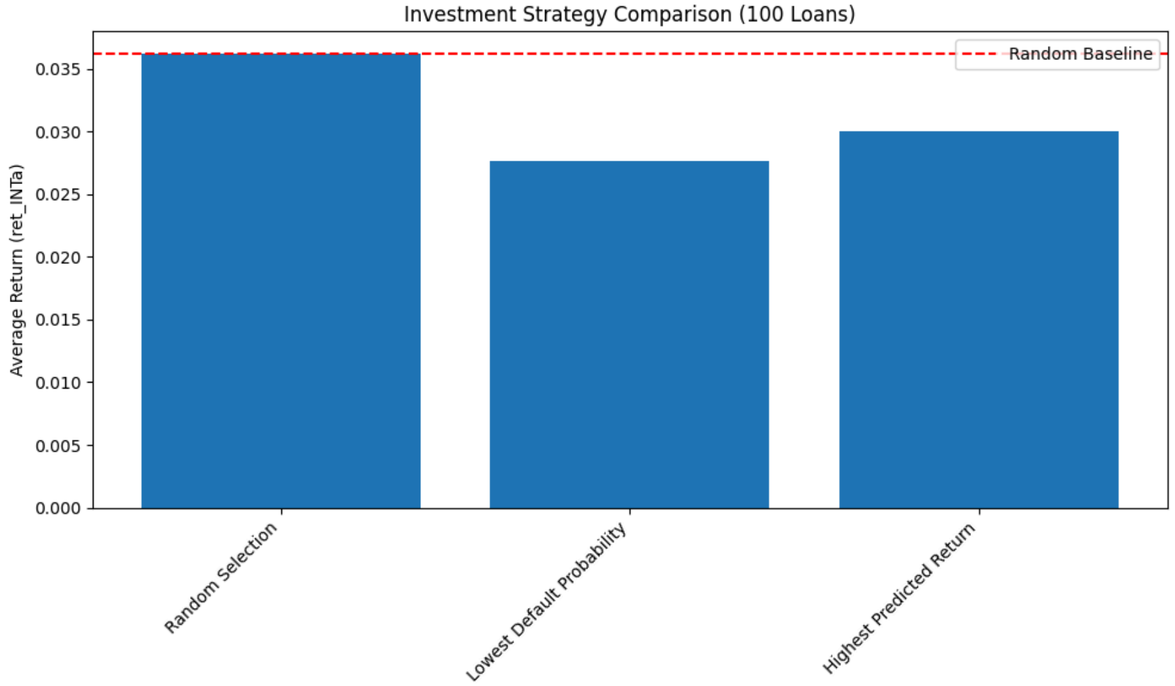
- ret_INTa: Return based on interest received
- ret_INTb: Return adjusted for principal loss
- ret_INTc: Return adjusted for both loss and recovery

Models Tested: Ridge Regression, Lasso Regression, ElasticNet

Features Used: 23 borrower + loan features (e.g., Interest Rate, Loan Amount, Loan Purpose, Employment Length, Revolving Credit Balance, etc.)

We Tested 3 Strategies (Using ret_INTa as our return metric)

Strategy	Description
Random Selection	Baseline: pick 100 loans at random
Lowest Default Probability	Choose 100 loans with safest scores from our best model
Highest Predicted Return	Use regression to predict returns & pick top 100



Random Selection performed best.

It gave a higher average return (~3.6%) than both:

- **The safest loans (low default risk)**
- **The best-looking loans (highest predicted return)**

Surprisingly, random loan selection outperformed our model-based strategies. This suggests our models may miss key return drivers, the market may efficiently price risk, or our test sample has unique traits. Jasmin should test on larger samples, explore hybrid models, and assess whether random selection's gains come with higher volatility.

CUSTOM STRATEGIES JASMIN COULD USE

Grade-Interest Rate Imbalance

Find loans where the interest rate is lower than expected for that grade

- Focused on Grades C, D, and E
- Selected loans where interest rate is 0.5 standard deviations below the grade's average
- These loans may be underpriced for their risk level

Low debt-to-income ratio and mortgage ownership

Choose loans from borrowers with low debt-to-income ratios and mortgage-backed home ownership

- Filtered bottom 30% for DTI
- Included only borrowers marked as 'MORTGAGE' owners
- These borrowers are likely financially stable

Temporal Pattern Arbitrage

Invest in loans issued at historically better-performing times

- Analyzed returns by month and day of week
- Chose loans issued in top 3 months and top 2 weekdays
- Added a secondary filter: Debt-to-Income within the top 40% lowest.

Results

Grade-Interest Rate Imbalance Strategy:

This strategy shows a modest but positive improvement of **6.39%** over random selection.

Low DTI + Mortgage Strategy

This strategy actually underperformed random selection by **7.89%**.

Temporal Pattern Arbitrage Strategy

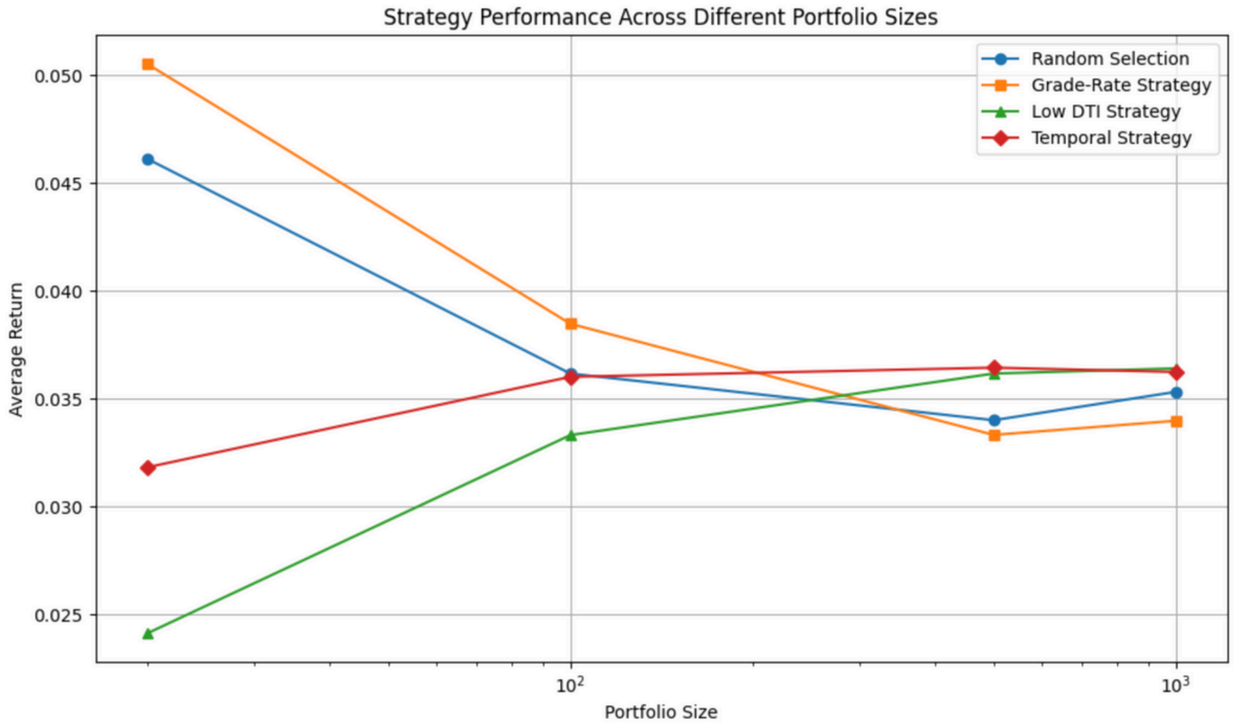
This is the most promising strategy, showing a **13.55%** improvement over random selection. months 2, 1, and 4 (and days 1 and 5) are particularly good

DO SMARTER STRATEGIES SCALE? PORTFOLIO SIZE VS. STRATEGY PERFORMANCE

What We Tested

We re-ran all four strategies (Random, Grade+Rate, Low debt to income ratio, Temporal) at different portfolio sizes: 20 loans, 100 loans, 500 loans, 1000 loans
We tracked:

- Average return using ret_INTa (interest earned divided by the amount invested)
- Relative improvement over the Random baseline



Key Insights

- **Random Strategy Performance:** Decreases from 0.046 (20 loans) to 0.034-0.036 (500+ loans), Returns stabilize with larger portfolio sizes
- **Grade-Rate Strategy:** Best at small portfolios (20 loans): 0.050539, Performance declines with portfolio size, Positive improvement only at smaller portfolios (9.54% at 20 loans)
- **Low DTI + Mortgage Strategy:** Poor initial performance (0.024 at 20 loans), Improves to 0.036 for 500-1000 loans, Significant underperformance at 20 loans (-47.74%), Becomes competitive at larger portfolios
- **Temporal Pattern Strategy:** Gradually improves with portfolio size, 20 loans: Underperforms significantly, 1000 loans: Modest positive return, Most consistent improvement at larger portfolios

Key Strategic Recommendations:

- Minimum viable portfolio: 100+ loans
- Optimal portfolio size: 500-1000 loans
- Temporal pattern strategy shows most promise for larger investments
- Significant strategy refinement needed for smaller portfolios

Investment strategy performance is highly dependent on portfolio size, with larger portfolios providing more stable and predictable returns.

CONCLUSION & NEXT STEPS
JASMIN'S SMARTER INVESTMENT JOURNEY

Final Recommendations to Jasmin

What We Learned

- Interest Rate is the strongest predictor of default.
- Logistic Regression + Downsampling improves risk identification.
- ret_INTa (interest-based return) is the most interpretable and predictable.

Strategy by Portfolio Size

- Small (≤100 loans) : Grade + Interest Rate Filter
- Medium–Large (500+) : Temporal Pattern Arbitrage (↑ 13.5%)

Key Takeaways

- Random selection surprisingly outperformed models → suggests market efficiency or missing signals.
- Custom strategies like temporal filters offer scalable improvements.
- Jasmin should test larger samples and balance return with risk.

LendingClub

THANK YOU