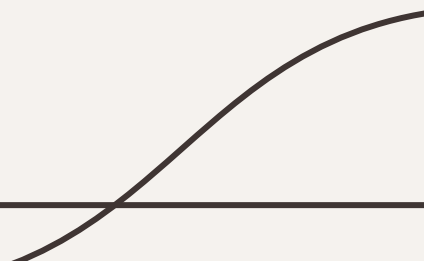


Data-Driven Investment Strategies

Kojo Dokyi
Srushti Kadam
Nivedha Gautham Raj



What is Signal Leakage and Why Does it Matter?



Signal Leakage

Using information that wouldn't actually be available to us in real-time



Results

Used information that we wouldn't be able to access otherwise and made our model too good at predicting
Ex. Loan recoveries



Outcome

The model predicted loan defaults with near-perfect accuracy.



Takeaway

This model is misleading since it includes information that will only be revealed in the future

Can Lending Club's Own Scores Predict Default?

Using Lending Club's scores like *loan grade* didn't work as well as hoped.



Original data had very few defaults making it hard for the model to predict.

Reduced number of samples for equal representation of default and non-default.



Makes model fairer in treating both default and non-default loans.

Make more realistic predictions about defaults leading to better-informed investments.



Predicting Defaults with Borrower Information

Features Used



Features available at time of loan application and represent borrower's self-reported information.



Features such as *annual income*, *home ownership* and *employment length*.

Modeling Approach

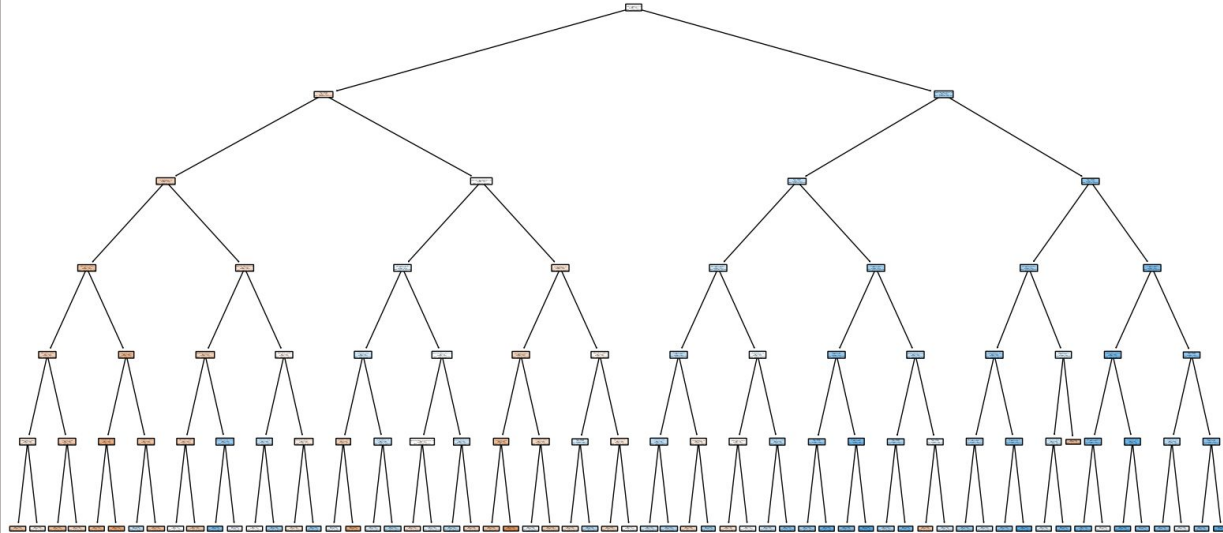
Built models only using borrower-supplied features to evaluate how well they can predict loan defaults.

Implications

- Isolates features within applicant's control and available publicly at time of application
- May have limited predictive power on their own

Top Rules to Predict Loan Default

Optimized Decision Tree (max_depth=6, class_weight='balanced')



Term of Loan > 60 Months

Borrowers with longer term of loan are at **higher risk** of default.

Revolving Utilization > 61.75 and Revolving Balance ≤ 2063.00

High risk when they are maxing out credit cards or struggling financially.

DTI ≤ 21.23 and Annual Income > 65046.50

Borrowers with low debt-to-income ratio and higher income are **low risk**.

Most Influential Features in Predicting Default

Higher Risk

Longer loan terms, higher debt-to-income ratio, recent credit inquiries, small business loans

Lower Risk

Higher income, larger revolving balances, mortgage owners

Takeaways

- Coefficients help understand why some loans are riskier than others.
- Provides good foundation for loan default prediction.

Feature	Coefficient
term_ 60 months	+0.3584
annual_inc	-0.1786
dti	+0.1650
revol_bal	-0.1547
inq_last_6mths	+0.1363
revol_util	+0.1041
purpose_small_business	+0.0953
home_ownership_MORTGAGE	-0.0838
home_ownership_RENT	+0.0833
open_acc	+0.0802

Selecting the Best Model for Predicting Defaults



Hypotheses

1. Longer term loans are riskier
2. Higher DTI increases default risk
3. Low-income borrowers more prone to default
4. Employment length and loan amount may influence risk.



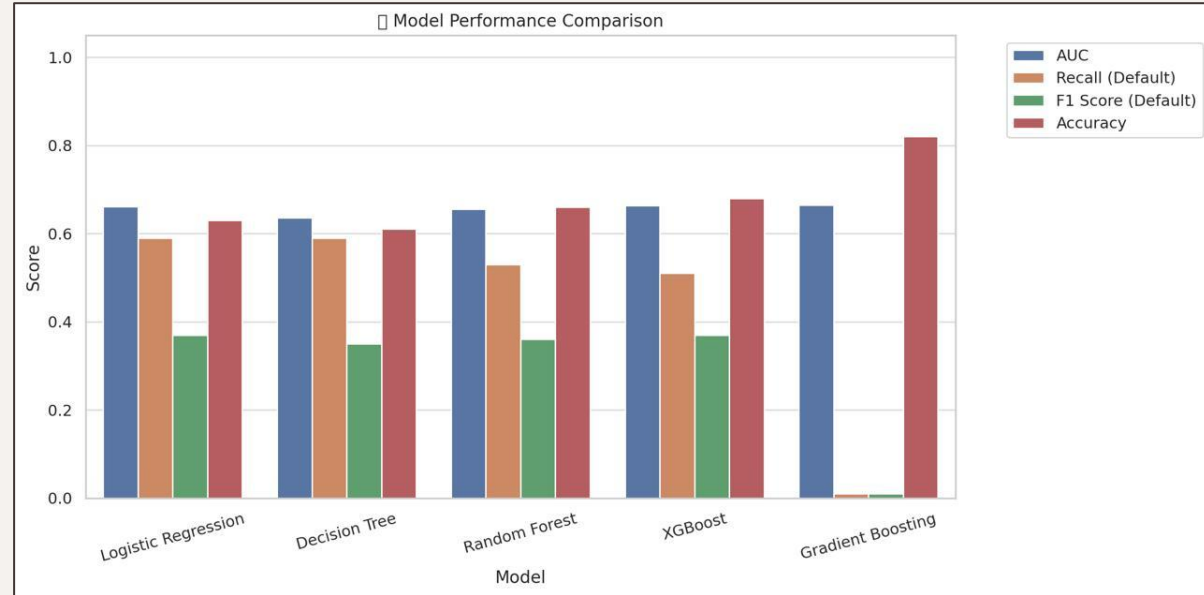
Observations

1. Validated
2. Validated
3. Validated
4. Partially validated – differs across models

Selecting Logistic Regression as the Best Model for Predicting Defaults

Key Takeaway

- Best combination of interpretability, stability and model performance
- **Interpretability:** easy understanding of how different variables affect default
- **Model performance:** high detection rate of defaults
- **Scalability:** easy to deploy to larger datasets if required



Model Comparison for Default Prediction

Key Takeaway

Evaluated three advanced models: Random Forest, XGBoost, and Gradient Boosting

- XGBoost achieved the best overall performance: Highest AUC (0.6631) and solid recall (0.51) for defaults
- Random Forest also strong: Balanced metrics with recall of 0.53
- Gradient Boosting had high accuracy (0.82) but failed to detect defaults (recall 0.01)

Recall for defaults is critical in lending accuracy alone is misleading

Recommendation: Use XGBoost for performance and Logistic Regression for transparency

Return Method Selection

Tested: ret_PESS, ret_OPT, ret_INTa, ret_INTb, ret_INTc

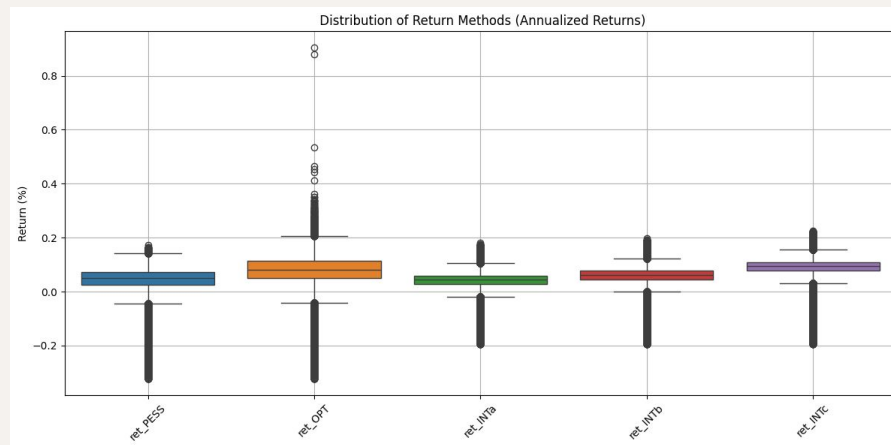
Chose: ret_INTc

Why?

- Highest avg. return (7.99%)
- Low downside risk
- Fewer outliers than ret_OPT
- More realistic than ret_PESS
- Assumes 0.5% reinvestment — investor-friendly

Conclusion:

ret_INTc = best balance of risk and reward
Used for all return predictions going forward



Return Prediction (ret_INTc)

XGBoost performed best using only application-time features, with the highest R^2 (0.33) and the lowest RMSE and MAE.

Tree-based models outperformed linear models, proving that early application data holds meaningful predictive power even without engineered features.

Model Performance Summary (on Test Set)

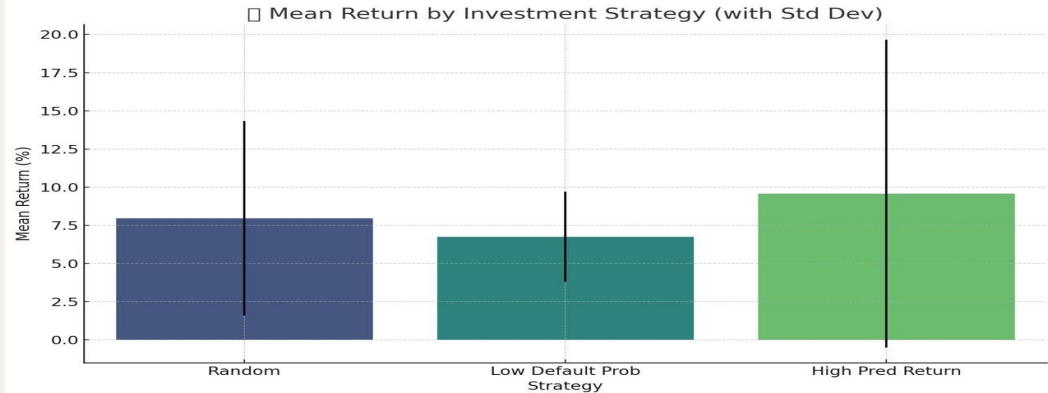
Model	RMSE	MAE	R^2 Score
Ridge	0.0540	0.0395	0.2646
Lasso	0.0544	0.0398	0.2555
ElasticNet	0.0542	0.0396	0.2603
Random Forest	0.0523	0.0336	0.3112
XGBoost	0.0516	0.0336	0.3297

Investment Strategy Comparison

We tested three strategies: random selection, lowest predicted default, and highest predicted return. The high return strategy gave the best average return (9.58%) but came with the highest default rate (29%). For lower risk, the low default strategy had only 5% defaults with more stable returns (6.76%), making it ideal for safer investing.

Summary Table

Strategy	Mean Return (%)	Std Dev (%)	Min Return (%)	Max Return (%)	Default Rate (%)
Random	7.97	6.37	-14.68	20.50	19.00
Low Default Prob	6.76	2.94	-11.78	8.98	5.00
High Pred Return	9.58	10.07	-19.21	22.09	29.00



Custom Strategies That Outperform

We built 4 custom strategies to find better loans.

Strategy 4 had the highest return (21.87%) with 0% defaults best overall.

Strategy 2 was the safest consistent 12.12% return and no volatility.

Strategy 1 found strong returns (16.44%) across interest rate bins.

Strategy 3 gave solid upside (17.28%) with light risk (3% defaults).

Strategy 4 is best for high return with low risk.

Strategy 2 is perfect for conservative investors.

Strategy 3 works for moderate risk takers.

Strategy 1 adds smart diversification.

Best move: lead with Strategy 4, mix with 1 or 2.

Strategy Comparison Summary

Strategy	Mean Return (%)	Std Dev (%)	Min Return (%)	Max Return (%)	Default Rate (%)
Best in Interest Rate Bin	16.44	4.09	9.98	22.66	0.0
Consistent High-Grade Loans	12.12	0.18	11.91	12.71	0.0
Hybrid (1 - Default) × Return	17.28	4.39	-1.93	22.15	3.0
Last-Minute High-Income Clean Borrowers	21.87	0.14	21.70	22.66	0.0

Strategy Strength Across Portfolio Sizes

- We tested all strategies at 20, 100, and 1000 loan sizes.
 - Hybrid and Last-Minute kept the highest returns (~20–22%) across all levels. Hybrid stayed at 0% defaults, even at 1000 loans.
 - Last-Minute had 1.1% defaults at scale but still strong.
 - Interest Rate Bin stayed consistent with ~15% return.
 - Consistent High-Grade was safest but returns dropped as size increased. For big portfolios, choose Hybrid or Interest Rate Bin.
- For smaller ones, use Last-Minute for gains or High-Grade for stability.

Strategy Scalability Summary

Strategy	Portfolio Size	Mean Return (%)	Std Dev (%)	Default Rate (%)
Best in Interest Rate Bin	20	14.65	3.08	0.0
Best in Interest Rate Bin	100	15.10	3.52	0.0
Best in Interest Rate Bin	1000	14.56	4.09	0.1
Consistent High-Grade Loans	20	12.39	0.15	0.0
Consistent High-Grade Loans	100	12.12	0.18	0.0
Consistent High-Grade Loans	1000	11.37	0.38	0.2
Hybrid (1 - Default) × Return	20	22.09	0.15	0.0
Hybrid (1 - Default) × Return	100	21.87	0.14	0.0
Hybrid (1 - Default) × Return	1000	20.20	0.99	0.0
Last-Minute High-Income Clean Borrowers	20	22.09	0.15	0.0
Last-Minute High-Income Clean Borrowers	100	21.87	0.14	0.0
Last-Minute High-Income Clean Borrowers	1000	20.21	0.98	1.1

Stability Over Time

The Hybrid strategy on 2014 data and tested it on 2015.

Return dropped from ~17–21% to **8.31%**.

Defaults jumped from ~3% to **31%**.

Volatility increased wider return range and higher std dev (10.22%).

Model patterns didn't hold across years.

Borrower behavior and LendingClub criteria likely changed.

Shows the need for time-aware validation and frequent retraining.

Conclusion:

Good models need to adapt one year of data isn't enough.