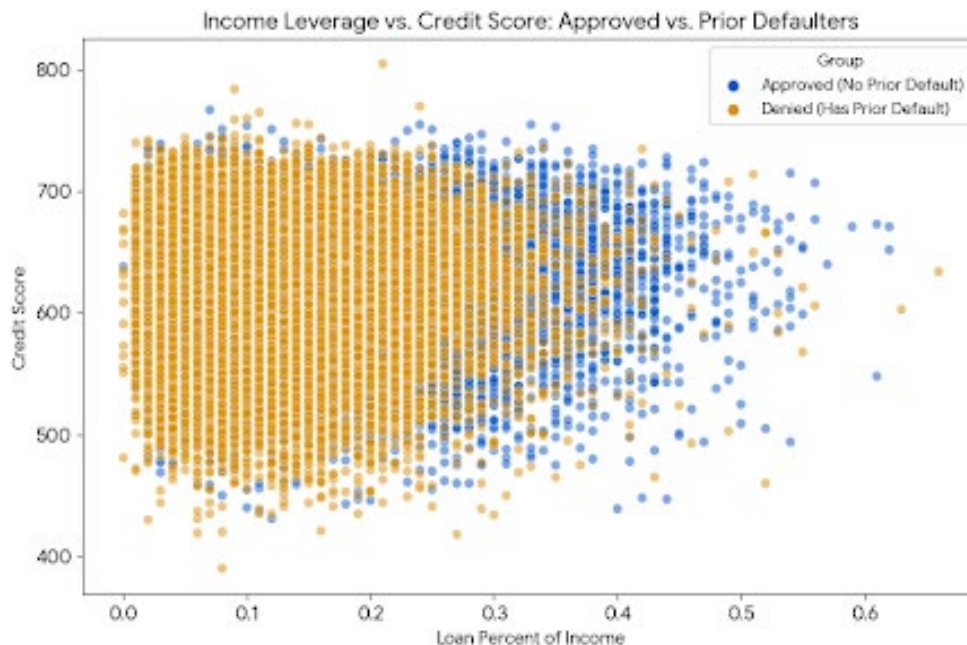


Base case #1: No automatic denial based on prior default history.

Our existing algorithmic approach to prescreening and denying applicants based on prior history of default is overly restrictive and limiting potential revenue. From a quantitative and business decision perspective this is a flaw in the decision rules applied to all applicants. Many of the 22,858 people denied for prior default have better credit scores and lower income-to-loan amount (leverage) ratios than the people currently getting approved. This represents a significant amount of lost revenue-generating loan activity.

Exactly 100% of applicants with a prior history of default are denied without human intervention or consideration of other most relevant criteria used to approve other applicants. We can solve this at the pre-screening stage using a predictive model. For applicants with a prior default, the model's target variable will output a direct recommendation for human intervention and manual review. When comparing applicants with and without a prior history of default, the credit score and income-to-loan amount leverage ratio are each among the top five most formative features of application approval data.

There is massive overlap between these two groups by credit score and loan amount as a percentage of income:



- **The "Approved" Group (No prior default):** Their average credit score is 631, and their median loan-to-income ratio (`loan_percent_income`) is 20%. Moreover, the probability of default for new applicants cannot be known beyond these criteria with current data at this prescreening stage.
- **The "Automatically Denied" Group (Prior default):** There are actually people in this group with credit scores soaring up to 805 (higher than the max score of the approved group, which topped out at 767!). Furthermore, ~50% of this excluded group is asking for a loan that is only 11% or less of their income, a much safer leverage ratio than the average approved person.

The current pre-screening algorithm has a technical flaw:

Because the historical data shows that exactly 22,858 people had a previous default and every single one of them was denied (`loan_status` = 0), the decision tree logic currently in place learned this as an absolute, unbreakable rule.

At a minimum, this represents a population of loan applicants we can target using predictive modeling to determine if an application should be recommended for human intervention (loan application review) for a final decision.

What would be a good target variable to find them?

Because the historical `loan_status` for this group is literally 100% "Denied" (0 variance), we cannot use it as a target variable to identify the good candidates. This forward-looking bias can be eliminated in the prescreening process by ignoring prior default and reserving it as a decision-making criterion for loan approval at the later human intervention and final decision stages.

To solve this and identify that highly-probable subset for reconsideration, we need to change our approach. Here are the best target variables to use:

1. A "Lookalike / Reconsideration Score" (Proxy Target)

Instead of predicting on the whole dataset, we filter out the prior defaulters entirely.

Step 1: Train a model (like Logistic Regression or a Random Forest) only on the people who had No Prior Defaults. Our target variable here is `loan_status`.

Step 2: Once the model learns what a "good" borrower looks like based on income leverage, credit score, etc., we pass the "Prior Default" group through this trained model.

The New Target Variable: The model will output a Predicted Probability (e.g., 0% to 100%). We can use this probability as our new continuous target variable. Anyone scoring over a certain threshold (say, >80% probability of being a good fit) represents our "Reconsideration Category."

2. Cluster Assignment (Unsupervised Learning)

If we don't want to use historical approvals at all, we can use a clustering algorithm (like K-Means) on `loan_percent_income`, `credit_score`, and `person_income`.

The New Target Variable: The Cluster ID.

We will likely find a "Low Risk, High Income, High Credit" cluster. We can then simply filter that specific cluster to see how many "Prior Defaulters" landed in it. Those are our premium reconsideration candidates.