**Model Report**

**Explorations relative to the target**

The dataset has 84.96% current loans and 15.04% defaults. This indicates a relatively healthy loan portfolio but also suggests a non-negligible risk rate that warrants a robust predictive model to mitigate potential defaults.

Loan Amount (loan\_amnt): Higher loan amounts may correlate with a higher likelihood of default. It would be beneficial to analyze the average loan amount for defaults versus current loans to identify any significant differences.

Interest Rate (int\_rate): There's a clear indication that higher interest rates are associated with higher default rates. Loans with a higher interest rate, such as 18.64%, might show a higher propensity for default compared to those with lower rates like 10.65%.

Loan Term (term): Longer terms, such as 60 months, exhibit a higher default rate compared to 36 months terms. This is crucial for risk assessment and loan pricing.

Loan Grade (grade): Lower grades indicate higher default rates, suggesting the effectiveness of the grading system in risk differentiation.

Based on the heatmap below, loan\_amnt, funded\_amnt, and funded\_amnt\_inv are highly correlated, which is logical because they all deal with the amount of the loan.

int\_rate and installment show a moderately positive correlation, suggesting that higher interest rates might be associated with larger installment payments.

total\_acc and open\_acc have a moderate positive correlation, likely because people with more open credit lines tend to have more accounts in total.

out\_prncp and out\_prncp\_inv have a perfect positive correlation, which indicates they are effectively representing the same information about the outstanding principal of a loan.

Negative correlations are also noticable like fico\_range\_low/fico\_range\_high and int\_rate, which could suggest that lower FICO scores might be associated with higher interest rates.

**A screen shot of a graph

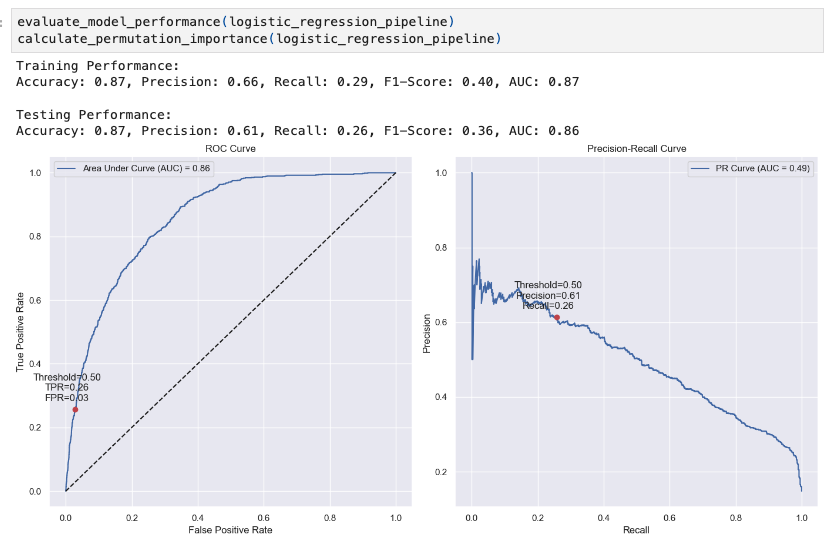
Description automatically generated**

**Anomaly Detection**

1. High Loan Amount with Low Income: An applicant with a significantly high loan amount but a reported annual income that seems disproportionately low could be an anomaly.
2. Zero or Extremely Low-Interest Rate:A loan with a 0% or unusually low-interest rate that does not align with market rates or the borrower's creditworthiness.
3. High Number of Delinquencies with a Good Grade: Records where a borrower has a high number of delinquencies but has been assigned a 'good' grade, such as A or B, might indicate a data entry error or misclassification.
4. Discrepancies in Loan Amount vs. Funded Amount: Significant discrepancies between loan\_amnt, funded\_amnt, and funded\_amnt\_inv could point to data integrity issues.
5. Inconsistent Credit History: Records with a high FICO score but with recent delinquencies or bankruptcies could be erroneous.
6. Suspicious Activity in Collections: Loans showing activity in collections (collections\_12\_mths\_ex\_med) but still marked as 'current' could be an anomaly or indicate recoveries on potentially written-off debts.
7. Mismatched Application Types: Loans that have an application\_type indicating individual application but have joint financial metrics (like combined annual income) may require further inspection.

**Model Comparison**

**Logistic regression**

****

**Neural Network** A screenshot of a graph

Description automatically generated

**Neural Network Optimized**

A screenshot of a graph

Description automatically generated

**Stacking Classifier**

A screenshot of a graph

Description automatically generated

The Stacking Classifier is the top-performing model, with a training AUC of 0.96 and testing AUC of 0.90. It surpasses logistic regression and neural network models in precision, recall, and F1-Score—a balanced metric for precision and recall, ideal for uneven class distributions or when false positives and negatives have similar costs. At the standard threshold of 0.5, it achieves a 41% true positive rate (TPR) and a 4% false positive rate (FPR). For optimal performance, adjusting the threshold to balance the F1-Score, considering the business implications of false positives and negatives, is recommended.

**Hyper parameter tuning**

**Random Forest Model**

I defined a parameter grid for the random forest with different numbers of estimators (n\_estimators), tree depths (max\_depth), minimum samples to split (min\_samples\_split), and minimum samples per leaf (min\_samples\_leaf). Then I performed a grid search using 3-fold cross-validation to find the best hyperparameters.

**Neural Network Model**

For the neural network model, I configured a parameter grid to tune the size of the hidden layer (hidden\_layer\_sizes), regularization term (alpha), and initial learning rate (learning\_rate\_init). A grid search with 2-fold cross-validation was employed for hyperparameter optimization.

**Stacking Classifier**

The stacking classifier was built using the best estimators from the previous steps, including a gradient boosting classifier, a random forest classifier, and a neural network, with their respective best-found hyperparameters.

**Best Model: Stacking Classifier**

The Stacking Classifier emerges as the superior model with robust metrics:

Accuracy: 92% during training and 88% in testing.

Precision: 84% for training, indicating trustworthiness.

Recall: 58% in training, suggesting it's reliable in identifying actual defaults.

F1-Score: 0.69 for training, signifying a well-rounded model.

AUC: 0.96 in training and 0.90 in testing, showcasing discriminative power.

Partial Dependence Plots (PDPs) as shown below provide nuanced insights:

Last Payment Amount (last\_pymnt\_amnt): The PDP suggests that higher last payment amounts generally decrease the model's predicted probability of default.

Interest Rate (int\_rate): As interest rates increase, so does the predicted probability of default, indicating that loans with higher interest rates carry a higher risk of default.

The PDPs for installment and annual income (annual\_inc) imply that as this increase, the likelihood of default slightly decreases, likely reflecting a borrower's increased capacity to manage and repay the loan. These interpretations from the PDPs enrich the understanding provided by variable importance metrics. This comprehensive performance and interpretability make the Stacking Classifier particularly advantageous for strategic financial decisions, where understanding the why behind predictions is as crucial as the predictions themselves.

A graph with numbers and a bar

Description automatically generated

A graph paper with lines

Description automatically generatedA screenshot of a graph

Description automatically generated

**Operational Strategy at 2% and at 5% FPR**

To achieve and maintain a 2% and 5% false positive rate (FPR), we would have to adjust the classification threshold of the model. This threshold determines the probability above which all loan applications are predicted to default. By default, this threshold is set at 0.5 (50%) for binary classifiers, meaning that if the model estimates the probability of default to be higher than 50%, it predicts a default. When we lower the threshold, we will catch more true positives but also more false positives and vice versa.

Operational Strategy at 2% FPR: First, find a threshold that results in a 2% FPR. This can be done by examining the ROC curve and identifying the point where the FPR is close to 2%.

Recall is likely to decrease because we will be identifying fewer loans as potential defaults.

Precision is likely to increase because a lower FPR means a higher proportion of the predicted defaults are defaults. Aiming for a 2% FPR means being very cautious about not mislabeling a good loan as a default. This will help maintain a good relationship with customers who are likely to pay back their loans. However, it also means that you might miss more defaulting loans, which could increase the risk of loss from defaults.

Operational Strategy at 5% FPR: Find a new threshold where the FPR is close to 5%.

Recall will be higher than at the 2% FPR threshold because we are allowing more loans to be classified as potential defaults.

Precision will be lower than at the 2% FPR threshold since we are allowing more false positives.

Overall Choosing a 5% FPR means we are willing to risk a few more incorrect predictions of default to catch more actual defaults. This strategy might help minimize credit loss from actual defaults but could also potentially alienate some customers who are wrongly classified as risky.

**Top 10 True Positives:**

For the top 10 true positives, the model was highly confident (probabilities close to or above 96%), suggesting that these loans had characteristics strongly indicative of default. Common traits among these might include:

High Loan Amounts: Large loans could contribute to higher risk, increasing the chance of default.

Higher Interest Rates: Typically associated with higher risk and potentially leading to greater difficulty in loan repayment.

Poor Credit History: If data on borrowers past financial behavior were available to the model, those with a history of late payments might be more likely to default.

These borrowers likely exhibited a combination of features that the model learned to associate with a high risk of defaulting.

**Top 10 False Positives:**

The false positives were cases where the model incorrectly predicted a default with high confidence. Possible reasons might include:

Outlier Behavior: These might be cases where the borrowers displayed risky behavior that is usually associated with defaulting but managed to stay current on their loan.

Data Noise: The model might be picking up on noise - idiosyncrasies in the data that aren't related to the outcome.

Overfitting to Anomalous Patterns: The model may have overfit to anomalous patterns in the training data that don't generalize well to the broader population.

**Top 10 False Negatives:**

For the false negatives, these were the loans that did default, but the model predicted they wouldn't, with very low confidence in default (probabilities around 3-4%). These might include:

Unexpected Defaults: Borrowers who, on paper, appeared to be low risk due to factors like a good credit score, moderate loan amounts, or solid income but ended up defaulting due to unforeseen circumstances (e.g., sudden financial hardship, job loss).

Insufficient Data: The model may lack key information that could have indicated a higher risk of default.

Subtle Indicators of Risk: The risk factors present may be subtler and not captured well by the model's learned patterns.

A screenshot of a computer

Description automatically generated

**Thank you for your time.**