Final Project – Loan Defaults

Can you predict loan default?

Anthony Ayala

Business Problem

Legacy Credit Union, a major financial institution in the United States, has faced persistent challenges within its loans and lending department. Despite its commitment to safety and support for not-for-profit associations, Legacy struggles with predicting loan defaults using Lending Club peer-to-peer lending data. As a new hire, I have been tasked to investigate the several factors that contribute to this issue, which are the following: inadequate explanations for predictions, uncertainty about predictors of fraud, and an overreliance on credit scores. The stakes are high: improving predictions directly impacts integrity, image, profitability, and potential business growth. By enhancing accuracy and minimizing errors, Legacy can attract new customers and gain deeper insights into loan default factors, ultimately preventing future defaults.

Methodology:

To make accurate predictions, the bread and butter is to train and evaluate different models based on performance metrics while managing the risk of overfitting. However, the real challenge lies in explaining this process effectively to colleagues and decision-makers. Here's the breakdown of how this process works:

1. Feature selection: choose relevant predictors through exploratory data analysis.
2. Partitioning the data: split the data into 80% training and 20% for testing.
3. Preprocessing: handle nulls values and scale numeric data.
4. Model training: train the model to recognize patterns crucial for predictions.
5. Parameter tuning: optimize the model's parameters for better performance.
6. Performance evaluation: assess various metrics to determine the best model.
7. Feature importance analysis: understand what factors contribute to predictions, especially for default and non-default loans.
8. Save predictions: save the predictions in a CSV file named df_submissions, indicating probability of default for each loan ID.

Key Findings

In the data I analyzed for Legacy prior to modeling, there were 25,300 loans that didn't default and 4,477 that did. It's important to note that Legacy only needs to focus on the 15% of loans that defaulted, aiming for accurate predictions in this subset. Initially, FICO/credit score didn't emerge as a strong predictor, nor did it rank in the top 10 for feature importance in our model. Instead, variables related to loan amount, funding, payments, and income emerged as significant predictors. Notably, loans predicted to default tended to have lower grades, higher late fees, smaller payments, and lower annual incomes.

Further analysis revealed additional predictive patterns, including factors like home ownership, delinquency history, and number of open accounts. Additionally, comparing true positive and false positive cases highlighted model bias, with the model placing different emphases on loan grade, amount, and payment method for each.

Model Performance and Interpretation:

We explored five different models and fine-tuned each one, ultimately selecting the stacking classifier as our best performer. This model combines outputs from multiple base models, enhancing predictive power. Its standout feature was its high AUC (Area Under the Curve), especially on test data, which is crucial for imbalanced datasets like ours, where non-defaults outnumber defaults. A high AUC indicates better discrimination between defaulted and non-defaulted loans, crucial for decision-making.

In addition to AUC, we focused on True Positive Rate (TPR) and False Positive Rate (FPR). TPR measures correctly predicted defaults among actual defaults, while FPR gauges incorrectly predicted non-defaults among actual non-defaults. We also examined the Precision-Recall Curve (PR Curve), which illustrates the trade-off between precision and recall at different thresholds. Precision ensures the accuracy of positive predictions (defaults), while recall reflects their completeness. For Legacy, precision is crucial for model correctness and aligns with the organization's goal of minimizing false positives (predicting non-defaulting loans as defaults), hence enhancing decision-making accuracy.

Operational False Positive Strategy:

Legacy seeks to minimize false loan default predictions while aiming for a 2% to 5% False Positive Rate (FPR). Our model achieves a precision of 62% and a recall of 48% at the 5% FPR level, meeting industry standards. At the stricter 2% FPR level, precision improves to 70%, emphasizing accurate identification of genuine default predictions, aligning with Legacy's goal of prioritizing fraud detection while minimizing errors in approving legitimate loans.

Actionable Recommendations:

I recommend that Legacy operates at a 2% False Positive Rate (FPR) to prioritize precision, crucial for our business problem. Assessing predictions for accuracy is essential, necessitating the recognition of model bias to limit both false negatives (type 2 errors) and false positives (type 1 errors). Notably, important predictors like late fees, last payment amount, loan grade, and loan amount can lead to incorrect predictions when their values are extreme and inconsistent with other variables. For instance, a loan predicted to default should not have contradictory features like a low grade but a high annual income. To handle false negatives, such as cases where abnormally high payments are incorrectly classified as defaults, further model tuning, more extensive training, or closer examination of false positives may be beneficial for improvement.

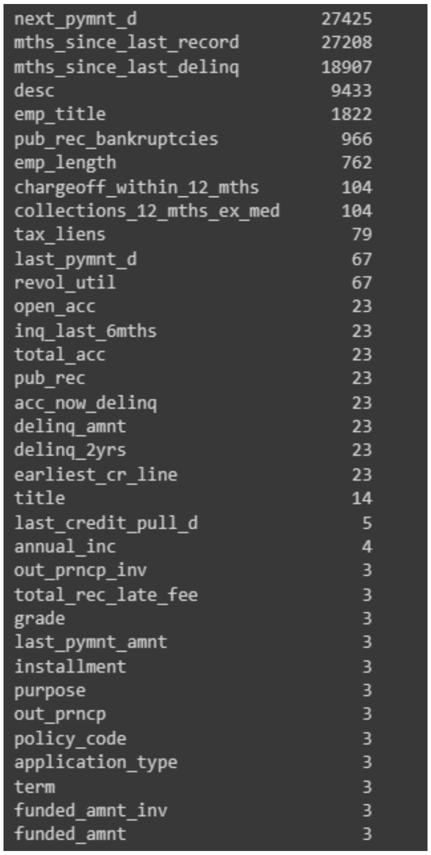
Exploratory Data Analysis

The Lending Club dataset comprises 29,777 rows and 52 columns, containing borrower information, loan details, and payment data. Of particular importance is the target variable, loan status, which we aim to predict. The dataset exhibits a clear class distribution: 25,300 rows representing loans that did not default, and 4,477 rows representing loans that did default, translating to an 85% non-default rate and a 15% default rate.

In terms of null analysis, the variable "loan status" stands out as it contains no missing values. However, several other variables, such as "next payment," "months since the last record," and "months since the last delinquency," have a significant number of null values. Addressing this issue is crucial, as simply filling in these missing values with the median or mean could introduce bias.

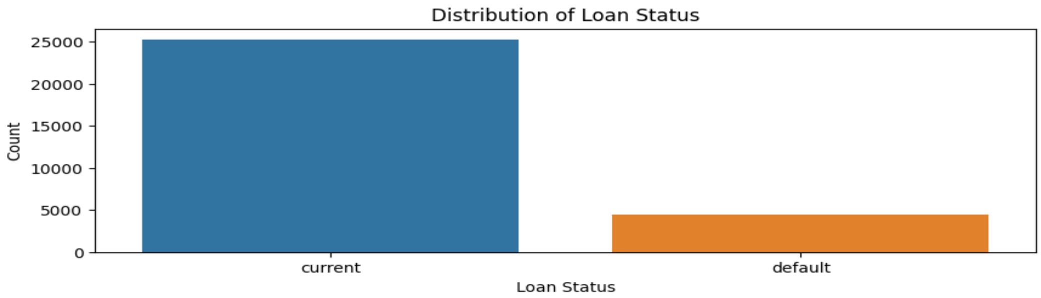
To handle this, I have opted to use the K-nearest neighbors (KNN) algorithm for numerical values. This approach identifies the nearest neighbors and fills in missing values based on their values, thus reducing bias. For categorical and non-numeric data, I have chosen to use the Simple Imputer, which fills in missing values with the most frequent value. This strategy aims to address extreme values by focusing on the most common and nearest data points, ensuring a more robust analysis.

Nulls



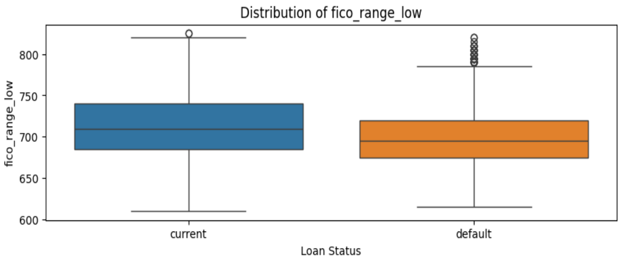
```
next_pymnt_d                 27425
mths_since_last_record       27208
mths_since_last_delinq       18907
desc                          9433
emp_title                     1822
pub_rec_bankruptcies           966
emp_length                     762
chargeoff_within_12_mths       104
collections_12_mths_ex_med     104
tax_liens                       79
last_pymnt_d                    67
revol_util                      67
open_acc                        23
inq_last_6mths                  23
total_acc                       23
pub_rec                         23
acc_now_delinq                  23
delinq_amnt                     23
delinq_2yrs                     23
earliest_cr_line                23
title                           14
last_credit_pull_d               5
annual_inc                       4
out_prncp_inv                    3
total_rec_late_fee               3
grade                            3
last_pymnt_amnt                  3
installment                      3
purpose                          3
out_prncp                        3
policy_code                      3
application_type                 3
term                             3
funded_amnt_inv                  3
funded_amnt                      3
```

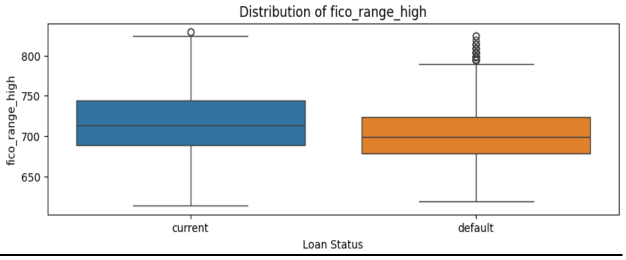
Distribution of Loan Status:

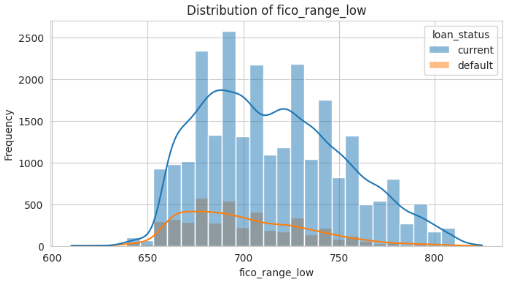


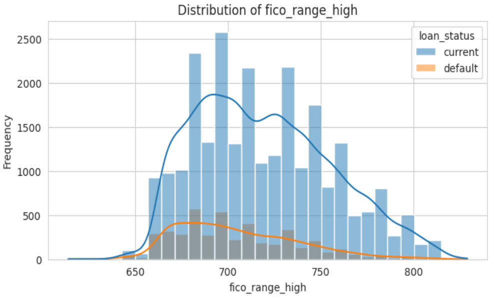
Before plotting variables against the target variable, I have decided to drop some variables that are irrelevant for prediction or that are biased such as zip code, URL, address state, job title.

FICO Ranges:





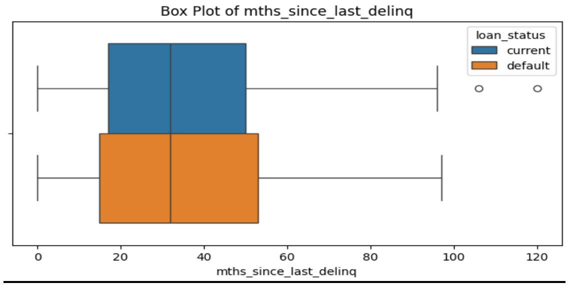


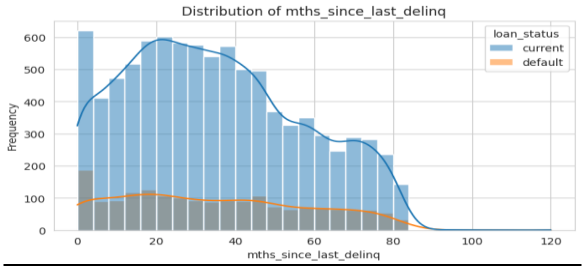


Legacy's assumption that defaulting loans have a direct relationship with FICO scores is understandable, given the conventional wisdom in the financial industry. However, upon analyzing the data, this assumption appears to be disproven. The data reveals that both low and high FICO scores do not clearly indicate a propensity for loan default within specific score ranges. While FICO scores may still be a contributing factor, they do not emerge as the strongest predictor or the sole determinant of loan default.

Of particular importance is the observation of outliers or anomalies in loans that defaulted despite having high credit scores, specifically above 770. This insight is significant as it prompts further investigation into the underlying reasons for these anomalies. Understanding why loans with seemingly excellent credit profiles still defaulted can provide valuable insights into factors beyond traditional credit scoring metrics that contribute to loan default risk. This underscores the complexity of credit risk assessment and highlights the necessity of comprehensive analysis beyond conventional assumptions.

Months Since Last Delinquent:

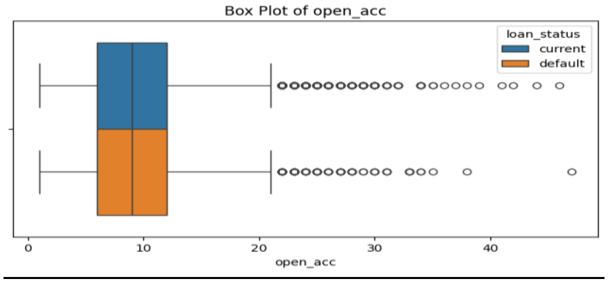


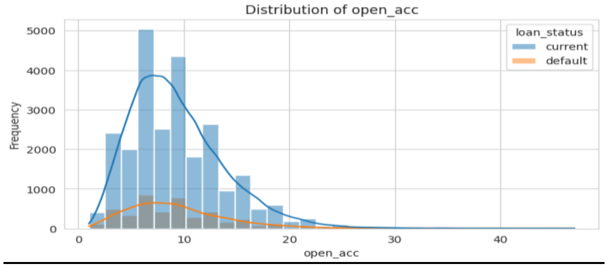


The label "Months since last delinquency" presents an intriguing perspective from a lender's standpoint. It reflects the duration since a borrower last failed to make timely payments, indicating a potential risk if they continue to struggle with payments. While there is a correlation between months of delinquency and loan status, it does not emerge as a robust predictor.

Moreover, to identify anomalies, it is noteworthy that only two records exist for loans that did not default, where the duration since the last overdue payment exceeded 100 months. This implies that these borrowers consistently made timely monthly payments, which is a remarkable observation given the usual variability in repayment behavior. Understanding such anomalies provides valuable insights into borrower behavior and repayment patterns, shedding light on exceptional cases that deviate from typical expectations.

Open Accounts:

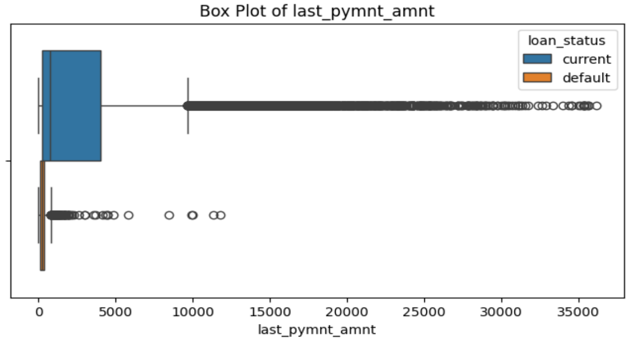


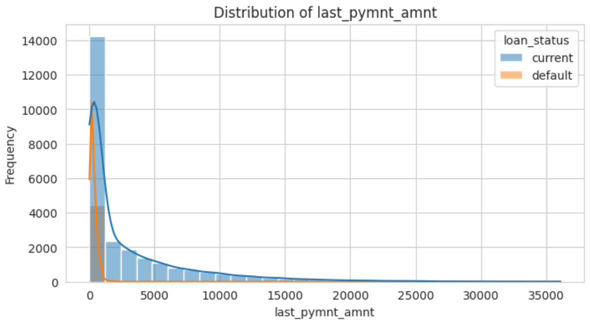


The number of open credit lines in a borrower's credit file is often perceived as an indicator of credit risk, with the assumption that having numerous open credit lines may negatively impact creditworthiness. However, upon examining the data, this assumption is not strongly supported. Despite the common belief, there is little discrepancy in the number of open credit lines between different loan statuses.

Furthermore, the presence of outliers or anomalies in the number of open accounts for both defaulted and non-defaulted loans suggests that there may be other factors at play beyond the sheer quantity of open credit lines. It is intriguing to note that a few borrowers with a high number of open credit lines did not default, indicating that additional variables may influence their default or non-default status. Unraveling the underlying factors contributing to these anomalies can provide valuable insights into the complex dynamics of credit risk assessment, challenging traditional assumptions and highlighting the need for a more nuanced understanding of borrower behavior.

Last Payment Amounts:

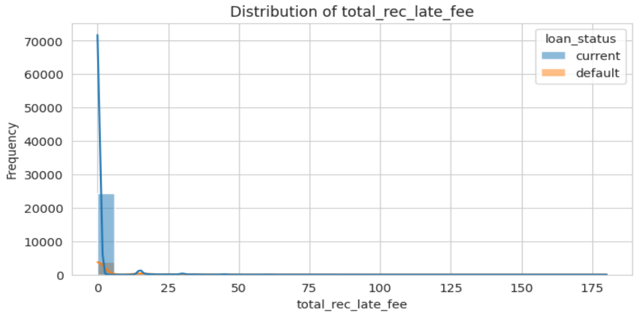


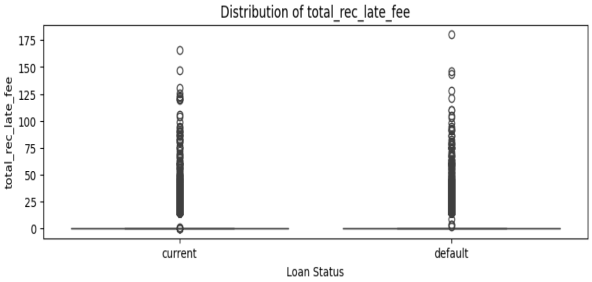


The last payment amount is often regarded as a key indicator of loan default, and this assumption is largely corroborated by the findings presented in both graphs. Specifically, instances where the last payment amount is $0 signify a clear indication of borrower default. Conversely, when the last payment amount exceeds $0, indicating that the borrower made at least the minimum payment, there is a notable absence of default.

In the box plot analysis, the presence of numerous anomalies or outliers further emphasizes this trend. Loans that did not default exhibit a pattern of substantial payments, suggesting a proactive approach to clearing debt and a strong commitment to avoiding default. Notably, making such large payments is atypical behavior, indicating either diligent saving by the borrower or a strategic decision to allocate funds towards loan repayment.

Total Late Fee:

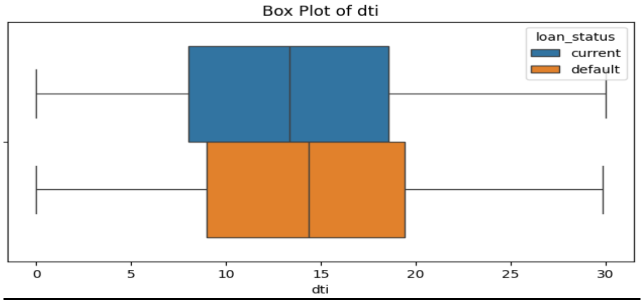


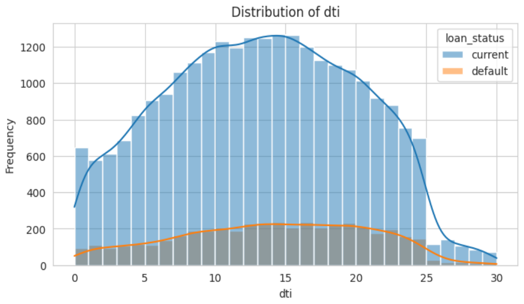


The total late fee presents a direct association with delinquency, as late payments typically incur penalties in the form of late fees. However, the graphs indicate that the amount of the late fee does not distinctly differentiate between loan statuses. Despite this observation, there remains a curiosity about whether this pattern will persist when modeling the data.

While the graphs do not reveal a significant disparity in loan status based on the late fee amount, delving into predictive modeling may provide further insights. It's possible that the impact of late fees on loan default risk becomes more apparent within the context of a comprehensive model. Exploring this aspect could uncover nuanced relationships between late fees and loan default, shedding light on the intricacies of credit risk assessment.

DTI (Borrower's Total Monthly Debt Payments / Borrower's Self-Reported Monthly Income):

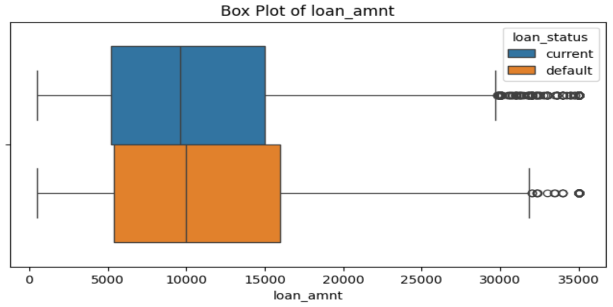


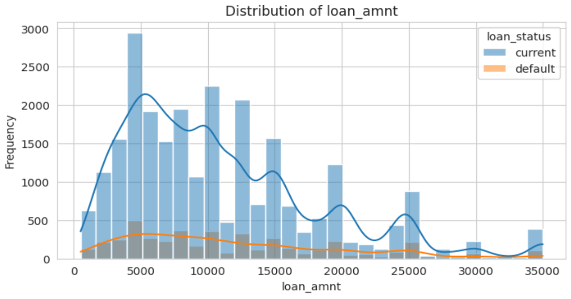


The debt-to-income (DTI) ratio is often considered a crucial factor in determining loan default risk, as it reflects a borrower's ability to manage debt relative to income. However, the graphs suggest that there is not a pronounced difference in loan default rates based on DTI. Additionally, it is notable that there are no outliers present in the DTI data.

Exploring the relationship between DTI and loan default in more depth, perhaps through predictive modeling or additional analysis, could offer valuable insights into the complex dynamics of credit risk assessment.

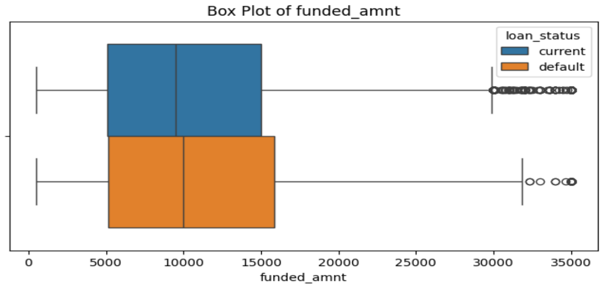
Loan Amount:

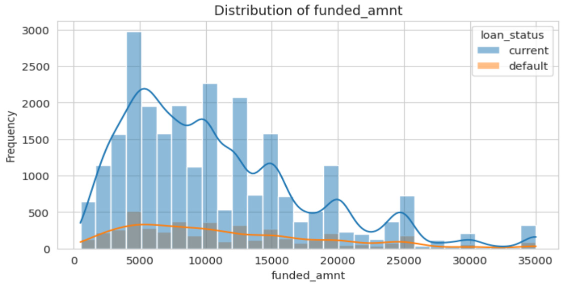




The loan amount does not seem to determine loan status as the plots show, but there is a large amount of outliers for both categories (loan status) especially in the higher end of a loan amount. It makes sense that loan amounts that are large could lead to defaulting on a loan as it could be too large to pay off or simply that the borrower was asking for too much.

Funded Amount:



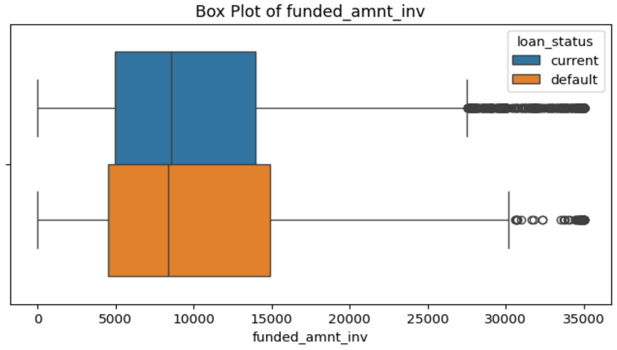


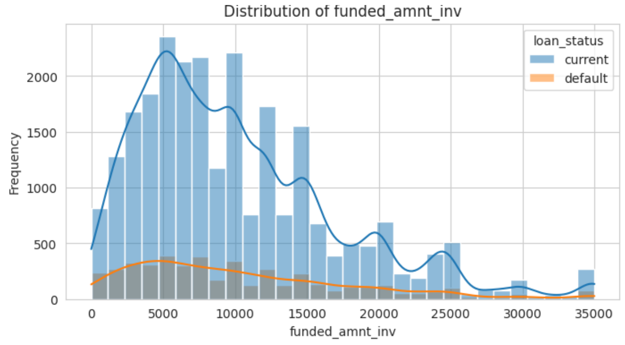
The loan amount does not appear to be a decisive factor in determining loan status, as suggested by the plots. However, it is noteworthy that a substantial number of outliers are present for both loan status categories, particularly at the higher end of the loan amount spectrum.

The observation of numerous outliers, especially in the higher loan amount range, raises intriguing questions about their significance in predicting loan default. It's reasonable to surmise that larger loan amounts could potentially increase the risk of default, either due to the burden of repayment or borrowers overextending themselves financially.

While the plots do not reveal a clear relationship between loan amount and loan status, the prevalence of outliers prompts further exploration. Analyzing these outliers in greater detail, perhaps through advanced statistical methods or predictive modeling, could unveil valuable insights into the nuanced interplay between loan amount and default risk.

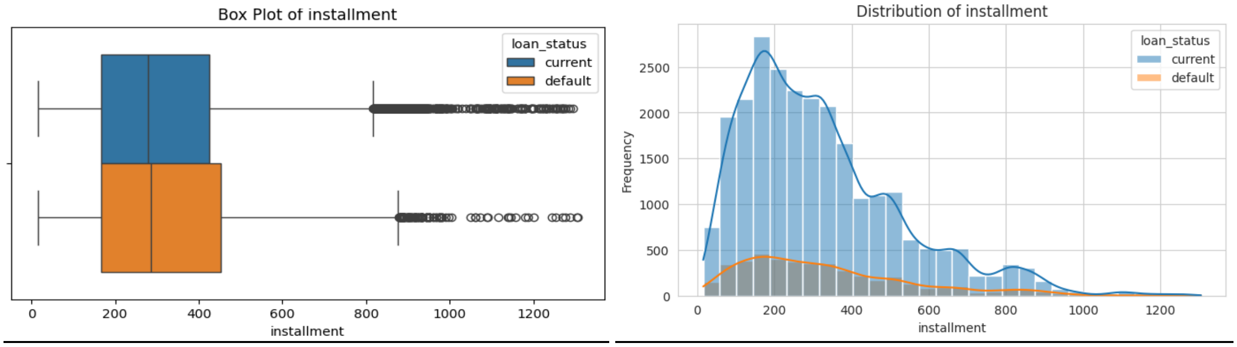
Funded Amount Investment:





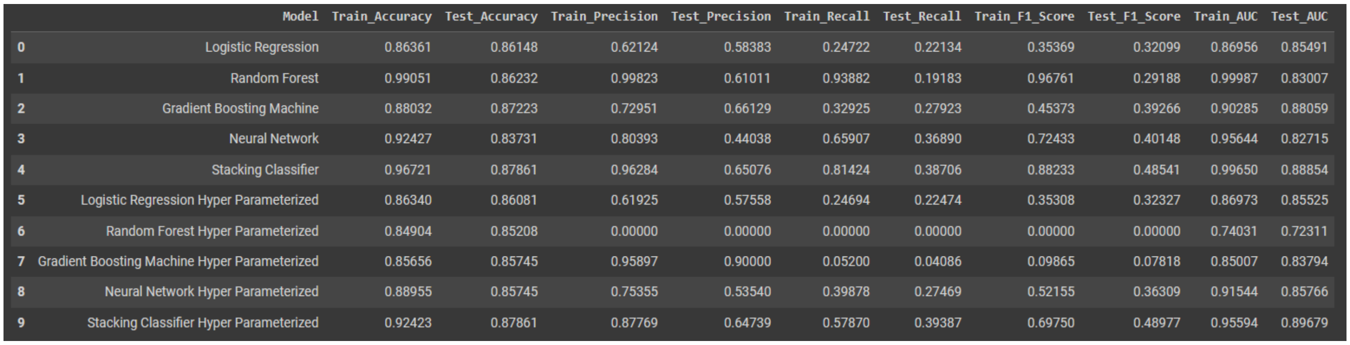
The funded amount investment doesn't strongly correlate with loan status, as shown in the plots. However, there are numerous outliers, particularly at higher investment levels. These outliers, especially in larger investments, raise questions about their impact on loan default. It's possible that excessive funding could increase default risk, either due to repayment challenges or borrowers receiving more than they can manage. Further analysis of these outliers could provide valuable insights into the relationship between funded amount and default risk.

Installment



While analyzing the plots, it becomes evident that the installment does not significantly influence loan status. However, there are notable outliers, particularly at higher installment amounts. The installment represents the borrower's monthly payment if the loan is originated, yet its magnitude doesn't seem to be a decisive factor in determining loan status. This observation underscores the importance of examining various factors beyond installment amounts when assessing loan default risk.

Model Performance and Metrics



	Model	Train_Accuracy	Test_Accuracy	Train_Precision	Test_Precision	Train_Recall	Test_Recall	Train_F1_Score	Test_F1_Score	Train_AUC	Test_AUC
0	Logistic Regression	0.86361	0.86148	0.62124	0.58383	0.24722	0.22134	0.35369	0.32099	0.86956	0.85491
1	Random Forest	0.99051	0.86232	0.99823	0.61011	0.93882	0.19183	0.96761	0.29188	0.99987	0.83007
2	Gradient Boosting Machine	0.88032	0.87223	0.72951	0.66129	0.32925	0.27923	0.45373	0.39266	0.90285	0.88059
3	Neural Network	0.92427	0.83731	0.80393	0.44038	0.65907	0.36890	0.72433	0.40148	0.95644	0.82715
4	Stacking Classifier	0.96721	0.87861	0.96284	0.65076	0.81424	0.38706	0.88233	0.48541	0.99650	0.88854
5	Logistic Regression Hyper Parameterized	0.86340	0.86081	0.61925	0.57558	0.24694	0.22474	0.35308	0.32327	0.86973	0.85525
6	Random Forest Hyper Parameterized	0.84904	0.85208	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.74031	0.72311
7	Gradient Boosting Machine Hyper Parameterized	0.85656	0.85745	0.95897	0.90000	0.05200	0.04086	0.09865	0.07818	0.85007	0.83794
8	Neural Network Hyper Parameterized	0.88955	0.85745	0.75355	0.53540	0.39878	0.27469	0.52155	0.36309	0.91544	0.85766
9	Stacking Classifier Hyper Parameterized	0.92423	0.87861	0.87769	0.64739	0.57870	0.39387	0.69750	0.48977	0.95594	0.89679

Here, I've outlined the metrics for each of the five models, along with their hyperparameter tuning, which I trained and tested. Initially, some models exhibited strong performance during training. However, upon testing, there was a noticeable decrease in performance. For instance, the Random Forest model achieved the highest training accuracy of 99.05%, but dropped to 86.23% during testing. This is a common occurrence as models are trained on existing data to make predictions on new, unseen data.

The breakdown of the metrics:

- Accuracy: Measures the proportion of correctly predicted loan defaults and non-defaults out of all predictions.

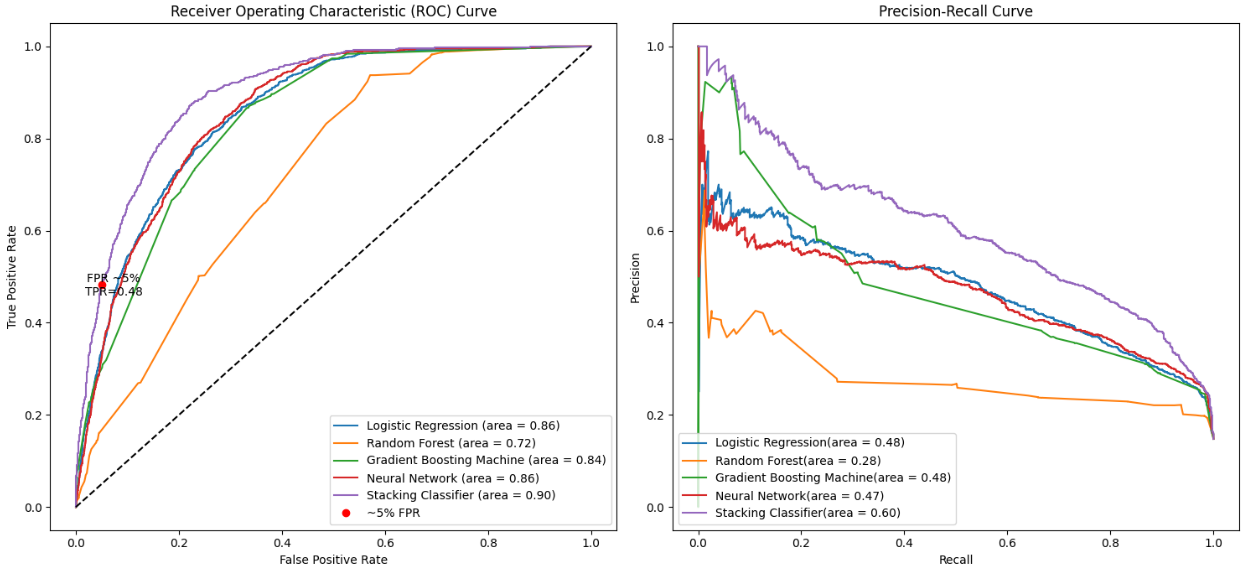
- Recall: Indicates the percentage of defaulted loans that were correctly identified as such among all actual defaults.

- Precision: Reflects the percentage of predicted defaults that actually turned out to be defaults among all predicted defaults.

- F1-Score: Balances precision and recall, with a perfect score of 1 indicating high precision and recall. It considers both false positives and false negatives, making it a robust measure.

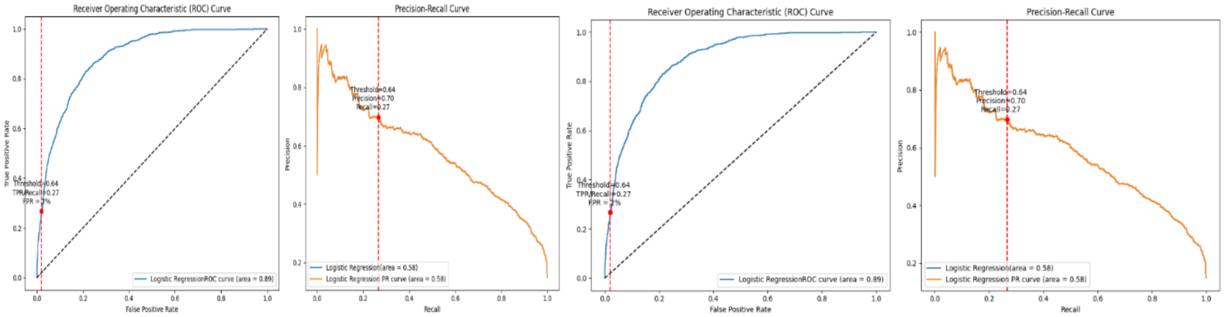
- AUC (Area under the curve): Evaluates a classifier's ability to distinguish between positive (defaults) and negative (non-defaults) classes. A higher AUC score signifies better discrimination between the two classes.

ROC Curves and Precision-Recall Curves



After analyzing the metrics and corresponding visuals, it's evident that the stacking classifier emerges as our best model. It boasts the highest AUC score, as well as impressive Precision and Recall Scores. These metrics, along with the F1-Score, are crucial for Legacy, as they aim to accurately identify loan defaults while minimizing false positives. The upcoming visual will detail the implementation strategy that Legacy plans to put into action based on these findings.

Operational Strategy 2% and 5% & Selecting a score 0.5:



As highlighted in the executive summary, Legacy aims to minimize false loan default predictions while maintaining a 2% to 5% False Positive Rate (FPR). Our best model achieves a precision of 62% and a recall of 48% at the 5% FPR level, meeting industry standards. At the stricter 2% FPR level, precision improves to 70%, ensuring accurate identification of genuine default predictions, aligning with Legacy's goal of prioritizing fraud detection while minimizing errors in approving legitimate loans.

Setting the FPR not only determines the percentage of incorrect default predictions but also establishes the minimum probability threshold for classifying loan defaults, which is 39%. The ROC curve illustrates the relationship between FPR and TPR (Recall), with TPR reaching 47% at the 5% FPR level, indicating a high level of accurate default predictions.

If Legacy opts for stricter criteria with a 2% FPR, precision increases to 70%, reflecting a heightened confidence in default predictions. However, there's a notable drop in recall to 27%, highlighting the trade-off between precision and recall. Given Legacy's mission, precision is prioritized to ensure the accuracy of default predictions, emphasizing the organization's confidence in its decisions.

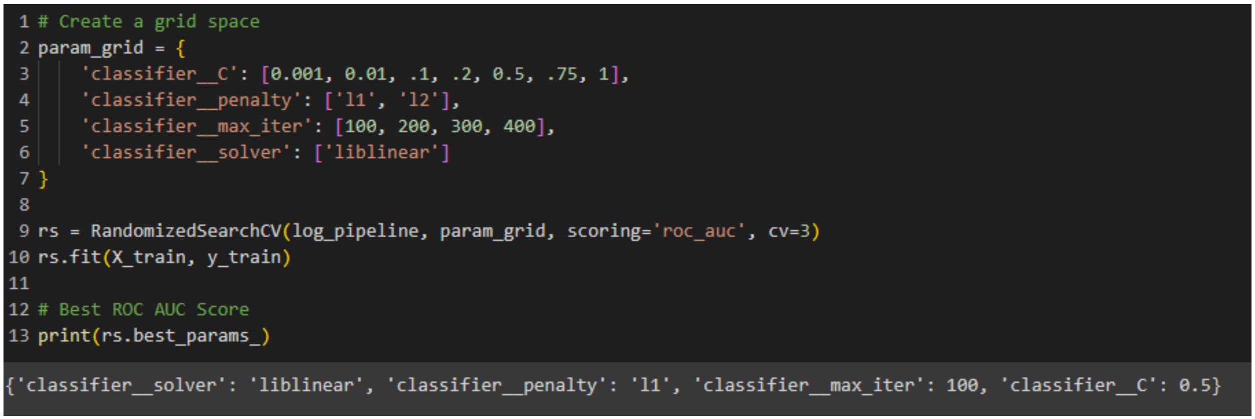
Hyper Parameters

My methodology for tuning the models involved optimizing each base model individually to ensure a fair assessment and select the best-performing model. I conducted 10 experiments, which proved to be sufficient for this purpose. To streamline the process and manage computational expenses, I opted for Randomized Search, striking a balance between search efficiency and tuning accuracy.

The primary optimization focus was on maximizing the AUC metric, which is more reliable than accuracy for imbalanced datasets like the one Legacy faces. A high AUC score ensures the classifier effectively distinguishes between classes (false positives and true positives), aligning with Legacy's requirements.

Each model underwent tuning for at least three parameters, except for the neural network due to its computational complexity. I maintained consistency in parameter selection between individual models and the stacking classifier, as the latter incorporates most of the former's components. Below, you'll find the models tested along with their best parameters.

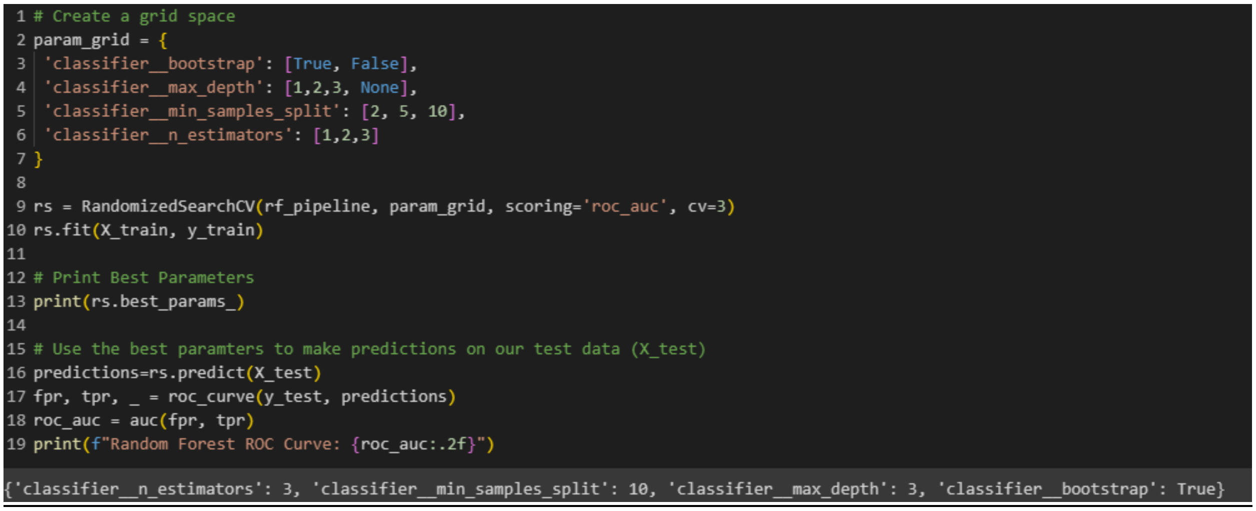
<u>Logistic Regression Tuning:</u>



```
# Create a grid space
param_grid = {
    'classifier__C': [0.001, 0.01, .1, .2, 0.5, .75, 1],
    'classifier__penalty': ['l1', 'l2'],
    'classifier__max_iter': [100, 200, 300, 400],
    'classifier__solver': ['liblinear']
}

rs = RandomizedSearchCV(log_pipeline, param_grid, scoring='roc_auc', cv=3)
rs.fit(X_train, y_train)

# Best ROC AUC Score
print(rs.best_params_)
```

```
{'classifier__solver': 'liblinear', 'classifier__penalty': 'l1', 'classifier__max_iter': 100, 'classifier__C': 0.5}
```

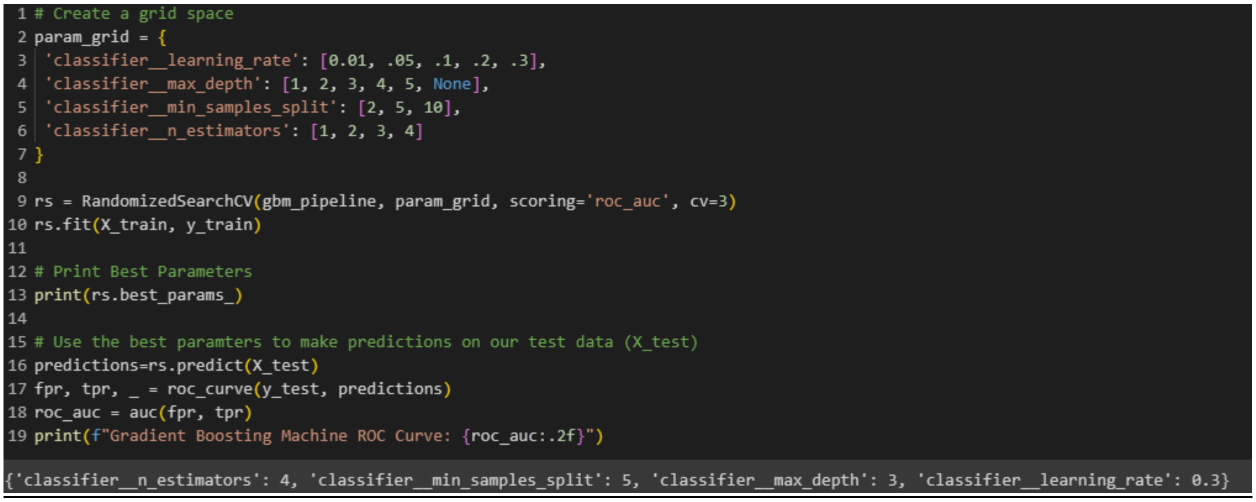
<u>Random Forest Tuning:</u>



```
# Create a grid space
param_grid = {
  'classifier__bootstrap': [True, False],
  'classifier__max_depth': [1,2,3, None],
  'classifier__min_samples_split': [2, 5, 10],
  'classifier__n_estimators': [1,2,3]
}

rs = RandomizedSearchCV(rf_pipeline, param_grid, scoring='roc_auc', cv=3)
rs.fit(X_train, y_train)

# Print Best Parameters
print(rs.best_params_)

# Use the best paramters to make predictions on our test data (X_test)
predictions=rs.predict(X_test)
fpr, tpr, _ = roc_curve(y_test, predictions)
roc_auc = auc(fpr, tpr)
print(f"Random Forest ROC Curve: {roc_auc:.2f}")
```

```
{'classifier__n_estimators': 3, 'classifier__min_samples_split': 10, 'classifier__max_depth': 3, 'classifier__bootstrap': True}
```

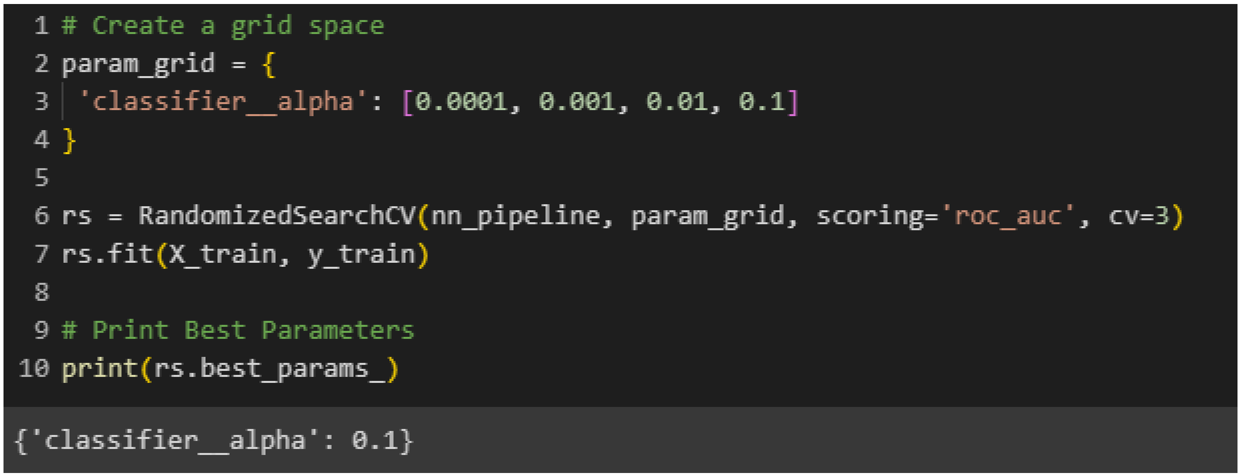
Gradient Boosting Machine Tuning:



```
# Create a grid space
param_grid = {
  'classifier__learning_rate': [0.01, .05, .1, .2, .3],
  'classifier__max_depth': [1, 2, 3, 4, 5, None],
  'classifier__min_samples_split': [2, 5, 10],
  'classifier__n_estimators': [1, 2, 3, 4]
}

rs = RandomizedSearchCV(gbm_pipeline, param_grid, scoring='roc_auc', cv=3)
rs.fit(X_train, y_train)

# Print Best Parameters
print(rs.best_params_)

# Use the best paramters to make predictions on our test data (X_test)
predictions=rs.predict(X_test)
fpr, tpr, _ = roc_curve(y_test, predictions)
roc_auc = auc(fpr, tpr)
print(f"Gradient Boosting Machine ROC Curve: {roc_auc:.2f}")
```

```
{'classifier__n_estimators': 4, 'classifier__min_samples_split': 5, 'classifier__max_depth': 3, 'classifier__learning_rate': 0.3}
```

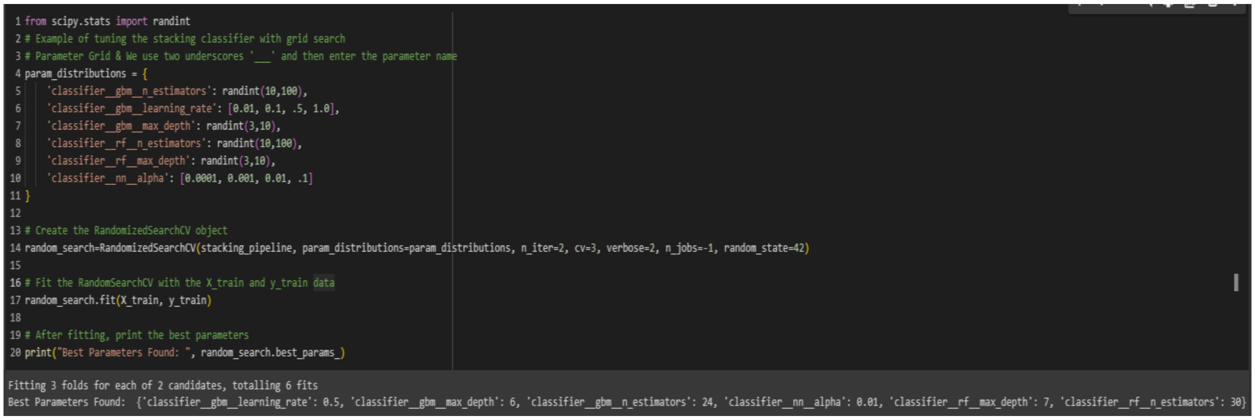
Neural Network Tuning:



```
# Create a grid space
param_grid = {
  'classifier__alpha': [0.0001, 0.001, 0.01, 0.1]
}

rs = RandomizedSearchCV(nn_pipeline, param_grid, scoring='roc_auc', cv=3)
rs.fit(X_train, y_train)

# Print Best Parameters
print(rs.best_params_)
```

```
{'classifier__alpha': 0.1}
```

Stacking Classifier Tuning:

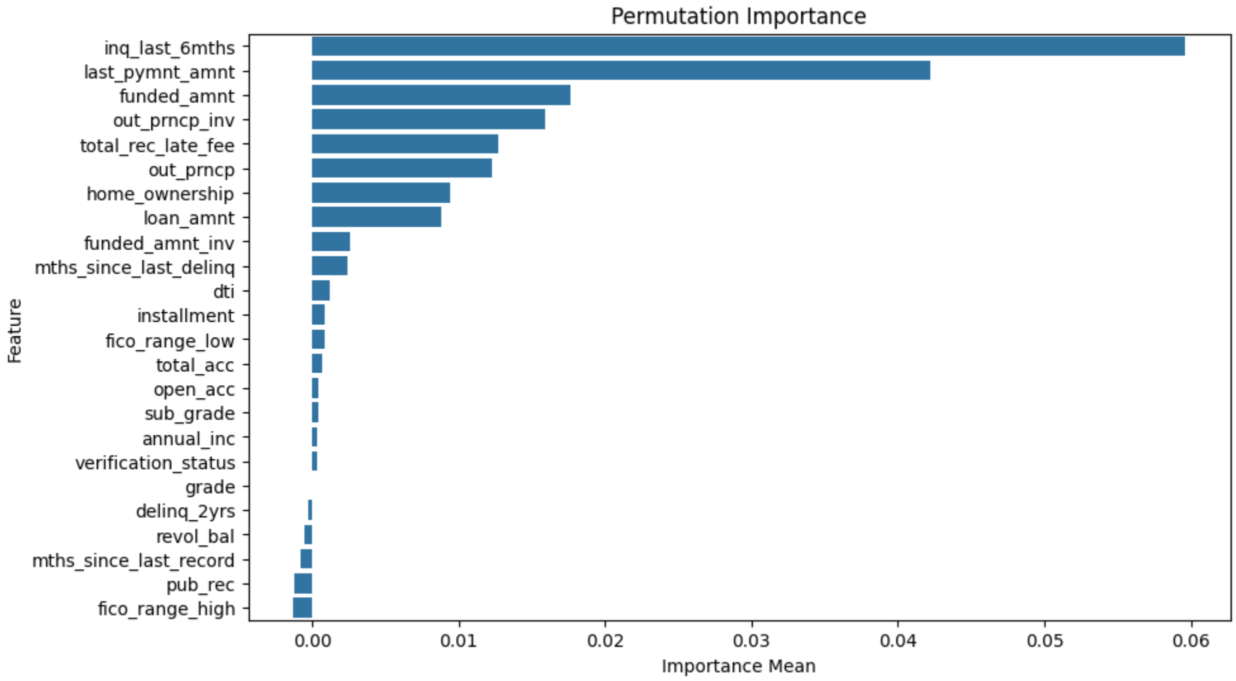


```python
from scipy.stats import randint
# Example of tuning the stacking classifier with grid search
# Parameter Grid & We use two underscores '__' and then enter the parameter name
param_distributions = {
    'classifier__gbm__n_estimators': randint(10,100),
    'classifier__gbm__learning_rate': [0.01, 0.1, .5, 1.0],
    'classifier__gbm__max_depth': randint(3,10),
    'classifier__rf__n_estimators': randint(10,100),
    'classifier__rf__max_depth': randint(3,10),
    'classifier__nn__alpha': [0.0001, 0.001, 0.01, .1]
}

# Create the RandomizedSearchCV object
random_search=RandomizedSearchCV(stacking_pipeline, param_distributions=param_distributions, n_iter=2, cv=3, verbose=2, n_jobs=-1, random_state=42)

# Fit the RandomSearchCV with the X_train and y_train data
random_search.fit(X_train, y_train)

# After fitting, print the best parameters
print("Best Parameters Found: ", random_search.best_params_)
```

```
Fitting 3 folds for each of 2 candidates, totalling 6 fits
Best Parameters Found:  {'classifier__gbm__learning_rate': 0.5, 'classifier__gbm__max_depth': 6, 'classifier__gbm__n_estimators': 24, 'classifier__nn__alpha': 0.01, 'classifier__rf__max_depth': 7, 'classifier__rf__n_estimators': 30}
```

Global Explanations of Best Model (Stacking Classifier)

Variable Importance



To determine the most influential variables in predicting loan default and non-default, I utilized permutation importance, a straightforward and easily implementable method. This technique evaluates a variable's impact by systematically altering its values and observing the resulting changes in model performance metrics like AUC, Accuracy, and error metrics such as MSE, MAPE, and MAD.

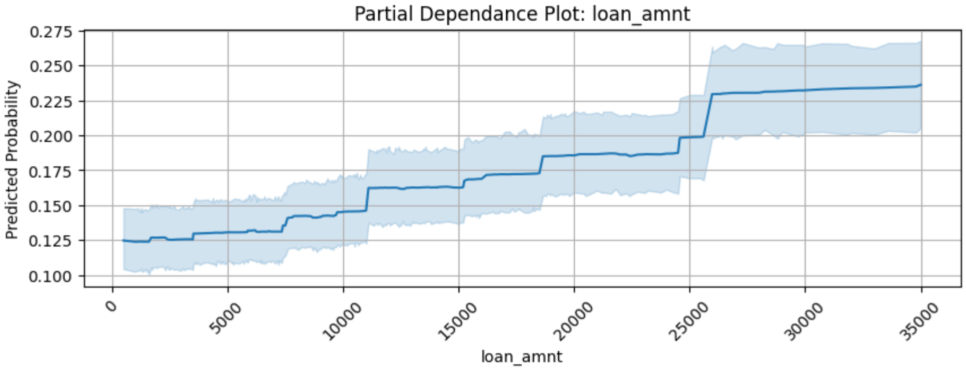
Our best model provided valuable insights by highlighting key variables, revealing potential oversights in our initial exploratory data analysis. The top ten variables include delinquency in the past 6 months, last payment amount, funded amount, remaining outstanding principal for the portion funded by investors, total late fee, outstanding remaining principal, type of home ownership, loan amount, funded amount by investors, and months since last delinquency. Interestingly, the remaining outstanding principal, not previously emphasized in our analysis, emerged as a significant predictor according to our model.

Furthermore, our analysis challenges the strong belief at Legacy regarding the impact of FICO/credit scores on predictions. Contrary to expectations, these scores appear to have minimal influence, prompting a reconsideration of their significance in loan default prediction.

Partial Dependance Plots of Top Variables:

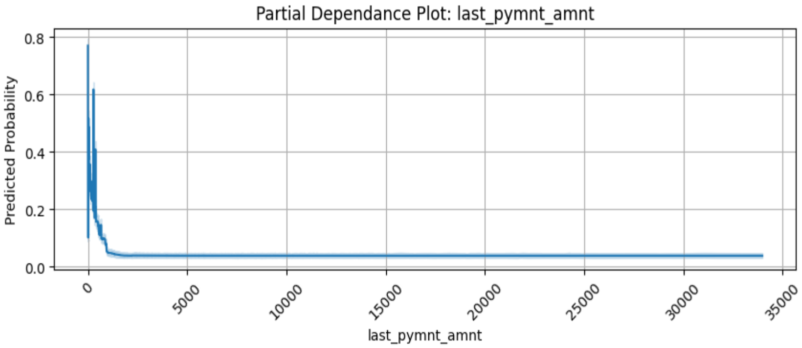
Partial dependence plots illustrate the average impact of specific input features on our predictions. They demonstrate the relationship between the target response and a selected set of input features, while averaging over the values of all other input features. Below, I'll display the most important variables identified through permutation importance and provide a brief explanation of their behavior in the plots.

Loan Amount:



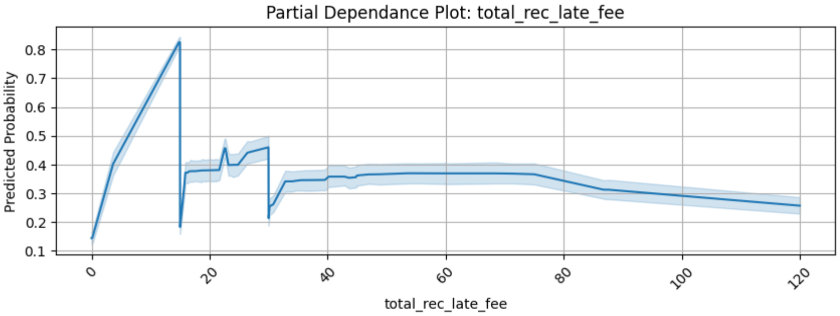
As loan amount rises, so does the likelihood of default. This insight underscores Legacy's understanding that loan amount reflects the borrower's requested funds. However, various additional factors related to loan amount can significantly influence the borrower's default risk.

Last Payment Amount:



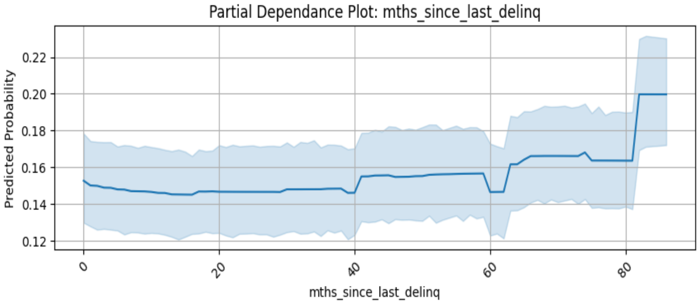
Our initial observation regarding the last payment amount remains valid: the likelihood of default is elevated when payments are $0 or very low. Conversely, as the last payment amount increases, the probability of default decreases significantly, with a slim or 0% chance observed in general. However, it's essential to note that this assumption will be integrated with other factors during the prediction process.

Total Late Fee:



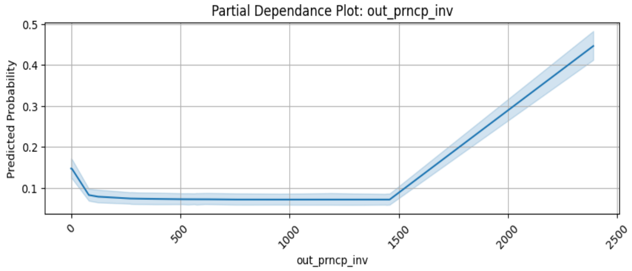
Total late fees typically indicate delinquency, often signaling a higher risk of defaulting on payments. However, an intriguing pattern emerges in our analysis. Initially, the probability of defaulting on a loan is notably elevated as the total late fee increases from $0 to approximately $18. This trend then reverses, with the probability decreasing before rising again within a different range. Eventually, it levels off within the $38 to $120 range. Despite fluctuations, the impact remains significant, with predicted probabilities ranging from 0.2 to 0.8 across various total late fee values.

Months Since Last Delinquent:



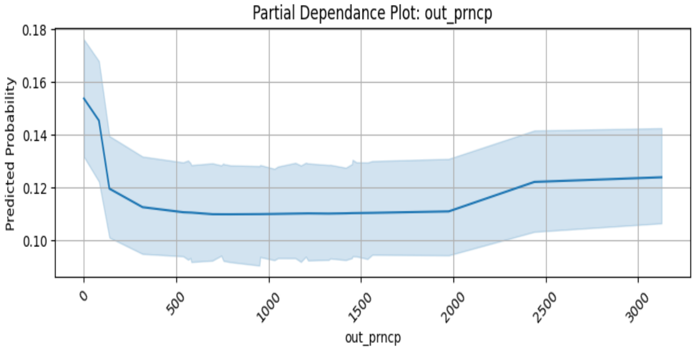
The conventional assumption was that the longer it has been since the last delinquency, the lower the probability of defaulting on a loan. However, our analysis reveals a different trend, potentially influenced by outliers observed during exploratory data analysis. While this challenges the assumption, what remains evident is that months since the last delinquency significantly impacts default likelihood. As the months increase, so do the chances of defaulting, suggesting a noteworthy relationship between these variables.

Remaining Outstanding Principle Investors:



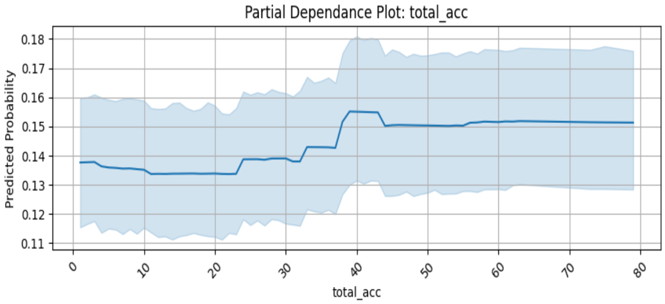
Surprisingly, this variable emerged as impactful in predicting loan defaults, contrary to expectations. We observe a significant increase in the probability of default when the remaining outstanding principal exceeds $1,500. This suggests that the remaining outstanding principal is a decisive factor in determining loan default likelihood, exerting a substantial influence on predictions.

Remaining Outstanding Principle:



Similarly, this variable was unexpected in terms of its impact on loan default probability, especially compared to the investor's version of the remaining outstanding balance. Contrary to expectations, when the balance is $0, there is a notable increase in the probability of loan default. However, as the balance increases, the predicted probability decreases. This comparison with the investor's version suggests a potential strong connection between loan funding and default risk, possibly indicating investors as key stakeholders.

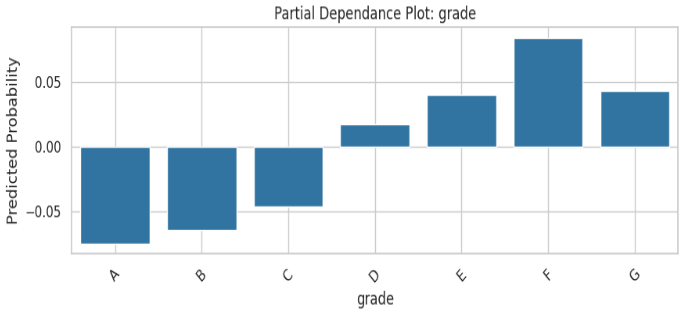
Total Accounts



I found total accounts to be an intriguing variable to examine, as the general assumption suggests that borrowers with more credit lines are riskier and tend to have lower credit scores.

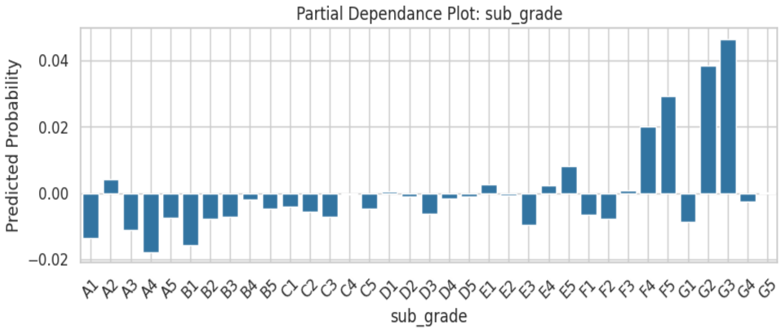
Surprisingly, we observe a clear association between the number of credit lines and loan default likelihood: as the number of credit lines increases, so does the probability of defaulting on a loan.

Loan Grade:



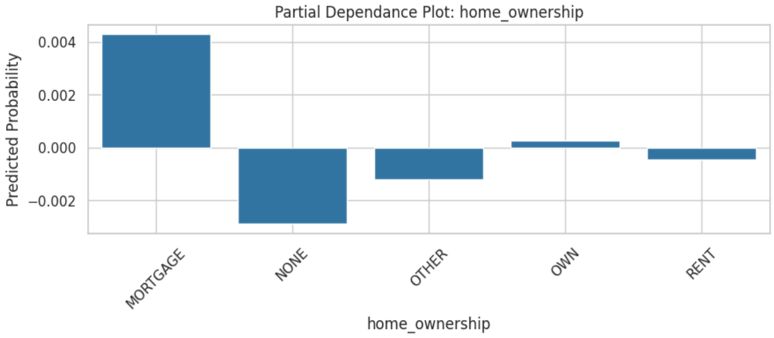
As a financial institution, it's crucial to assess the grade of distributed loans. Our analysis aligns strongly with the assumption that higher-quality graded loans have lower, if not decreased, chances of default. Conversely, lower-grade loans exhibit higher probabilities of default.

Loan Sub-Grade:



Breaking down loan grades into sub-grades provides valuable insights into their impact on the predicted probability of loan default. For instance, sub-grades within the A category generally decrease the likelihood of default, except for A2, which shows an increase in default probability. Identifying such nuances is crucial for understanding how variables contribute to the predicted probability of default, especially when analyzing true positives, false positives, and false negatives later on.

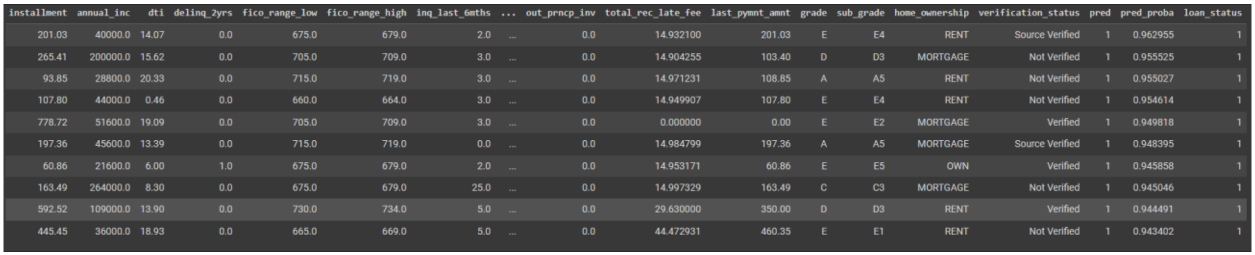
Home ownership:



The type of home ownership is a significant variable of interest, shedding light on the financial burden borrowers may face. Borrowers with a mortgage, for example, may experience increased default risks due to the added financial strain. It's noteworthy that ownership shows a slight increase in default, suggesting a need for further investigation into additional variables associated with this home ownership category to better understand the reasons behind borrower defaults.

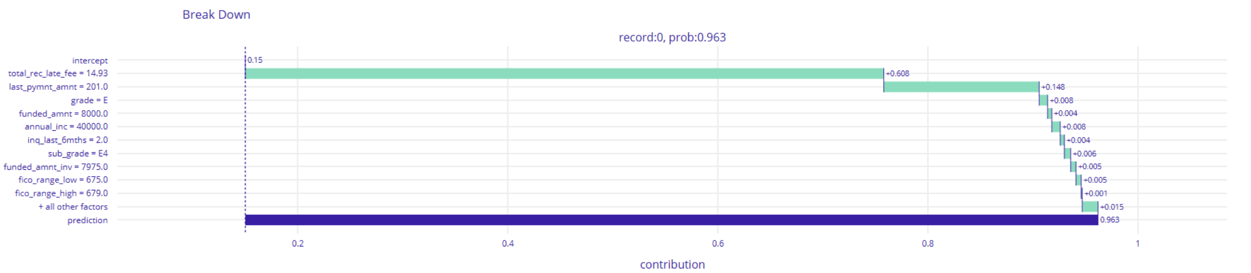
Local Explanations of Top 10 True Positives, False Positives, and False Negatives

<u>True Positives:</u>

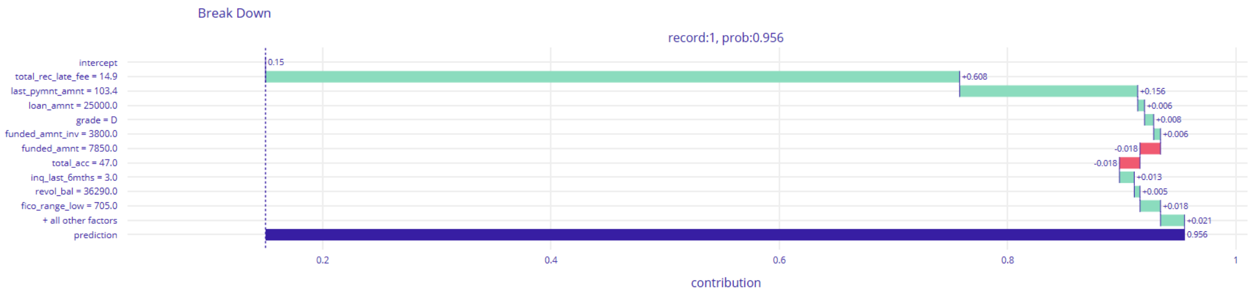


installment	annual_inc	dti	delinq_2yrs	fico_range_low	fico_range_high	inq_last_6mths	...	out_prncp_inv	total_rec_late_fee	last_pymnt_amnt	grade	sub_grade	home_ownership	verification_status	pred	pred_proba	loan_status
201.03	40000.0	14.07	0.0	675.0	679.0	2.0	...	0.0	14.932100	201.03	E	E4	RENT	Source Verified	1	0.962955	1
265.41	200000.0	15.62	0.0	705.0	709.0	3.0	...	0.0	14.904255	103.40	D	D3	MORTGAGE	Not Verified	1	0.955525	1
93.85	28800.0	20.33	0.0	715.0	719.0	3.0	...	0.0	14.971231	108.85	A	A5	RENT	Not Verified	1	0.955027	1
107.80	44000.0	0.46	0.0	660.0	664.0	3.0	...	0.0	14.949907	107.80	E	E4	RENT	Not Verified	1	0.954614	1
778.72	51600.0	19.09	0.0	705.0	709.0	3.0	...	0.0	0.000000	0.00	E	E2	MORTGAGE	Verified	1	0.949818	1
197.36	45600.0	13.39	0.0	715.0	719.0	0.0	...	0.0	14.984799	197.36	A	A5	MORTGAGE	Source Verified	1	0.948395	1
60.86	21600.0	6.00	1.0	675.0	679.0	2.0	...	0.0	14.953171	60.86	E	E5	OWN	Verified	1	0.945858	1
163.49	264000.0	8.30	0.0	675.0	679.0	25.0	...	0.0	14.997329	163.49	C	C3	MORTGAGE	Not Verified	1	0.945046	1
592.52	109000.0	13.90	0.0	730.0	734.0	5.0	...	0.0	29.630000	350.00	D	D3	RENT	Verified	1	0.944491	1
445.45	36000.0	18.93	0.0	665.0	669.0	5.0	...	0.0	44.472931	460.35	E	E1	RENT	Not Verified	1	0.943402	1

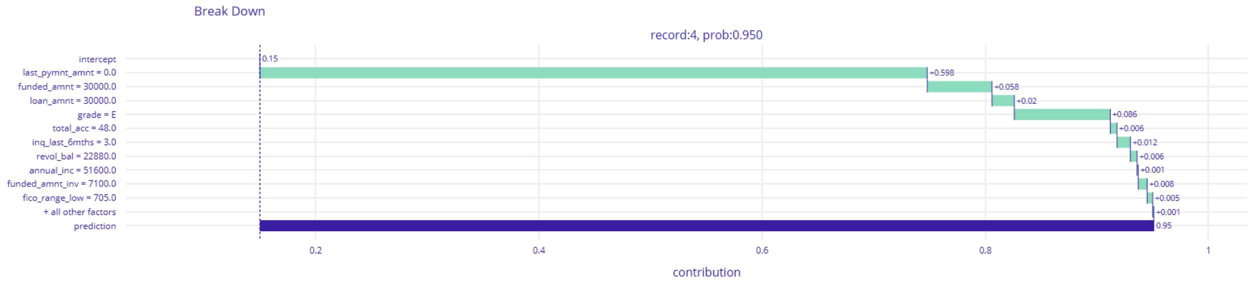
This isn't a complete table of the top 10 correctly predicted defaulted loans, but we can observe similarities in variables such as FICO range, total late fee, home ownership, and delinquency in the last 6 months. The graphs below will further dissect these records and elucidate their contributions to the probability of default.

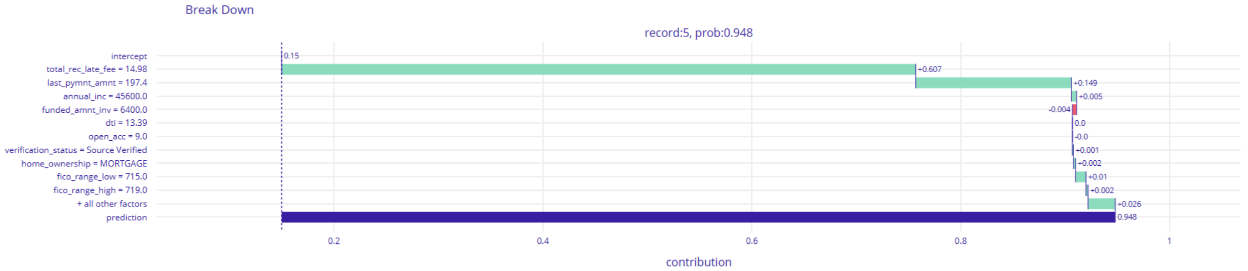


Total Late Fee emerges as the primary contributor, with a last payment of approximately $200 and an E loan grade indicating higher risk. Other factors contribute minimally to the overall predicted probability of default. This instance underscores the significant impact of factors such as poor loan grade, low payments, and late fees on the likelihood of default.



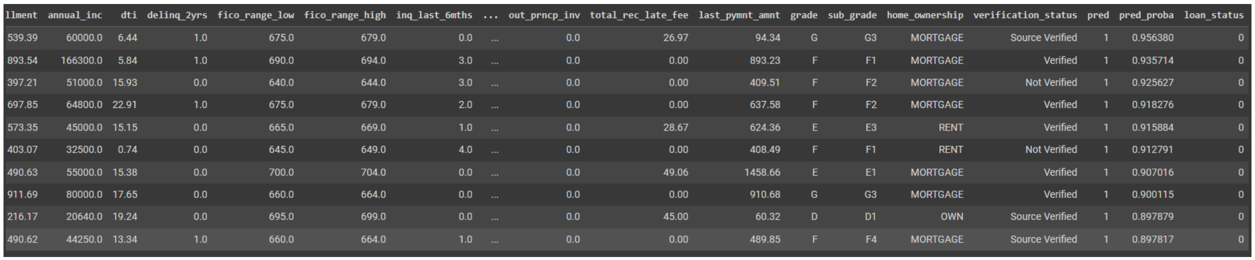
The same narrative applies here, but with notable differences. Total funded amount and 47 total accounts surprisingly decreased the probability of default. While this may seem counterintuitive, our model underscores the significance of other factors such as FICO range and revolving balance, which likely exert a stronger influence on the predicted probability of default.





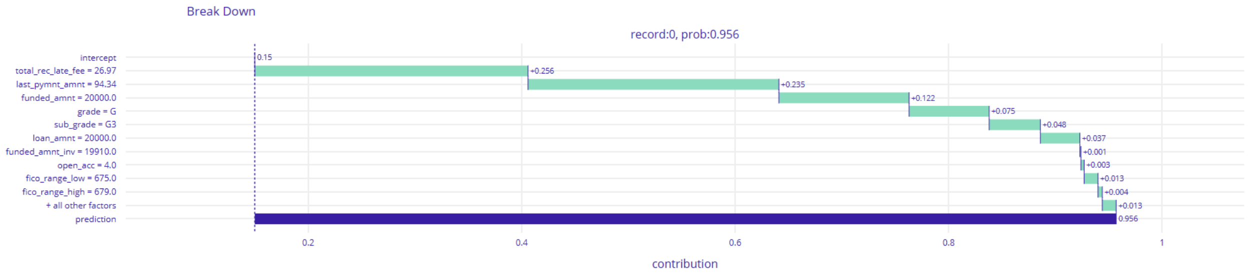
In reviewing the top 10 cases, we observe a sample breakdown that illustrates common drivers of predicted probabilities of default. These include borrowers with poor loan grades, low payments, minimal payment fees, high funded amounts, and insufficient annual income. These shared characteristics significantly influence the likelihood of default as revealed by our analysis.

False Positives:

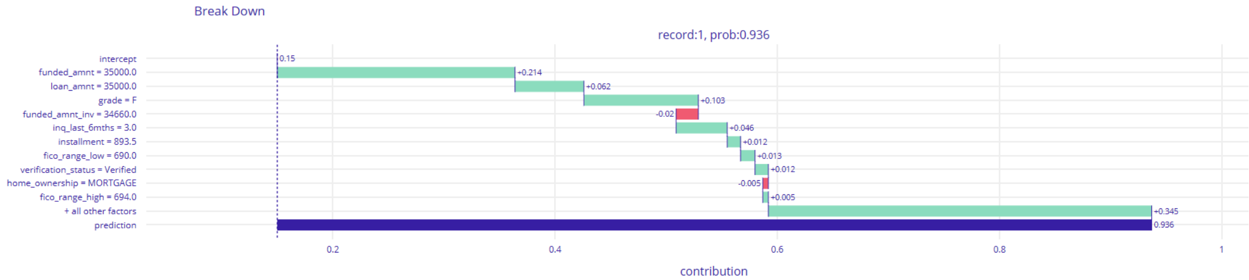


llment	annual_inc	dti	delinq_2yrs	fico_range_low	fico_range_high	inq_last_6mths	...	out_prncp_inv	total_rec_late_fee	last_pymnt_amnt	grade	sub_grade	home_ownership	verification_status	pred	pred_proba	loan_status
539.39	60000.0	6.44	1.0	675.0	679.0	0.0	...	0.0	26.97	94.34	G	G3	MORTGAGE	Source Verified	1	0.956380	0
893.54	166300.0	5.84	1.0	690.0	694.0	3.0	...	0.0	0.00	893.23	F	F1	MORTGAGE	Verified	1	0.935714	0
397.21	51000.0	15.93	0.0	640.0	644.0	3.0	...	0.0	0.00	409.51	F	F2	MORTGAGE	Not Verified	1	0.925627	0
697.85	64800.0	22.91	1.0	675.0	679.0	2.0	...	0.0	0.00	637.58	F	F2	MORTGAGE	Verified	1	0.918276	0
573.35	45000.0	15.15	0.0	665.0	669.0	1.0	...	0.0	28.67	624.36	E	E3	RENT	Verified	1	0.915884	0
403.07	32500.0	0.74	0.0	645.0	649.0	4.0	...	0.0	0.00	408.49	F	F1	RENT	Not Verified	1	0.912791	0
490.63	55000.0	15.38	0.0	700.0	704.0	0.0	...	0.0	49.06	1458.66	E	E1	MORTGAGE	Verified	1	0.907016	0
911.69	80000.0	17.65	0.0	660.0	664.0	0.0	...	0.0	0.00	910.68	G	G3	MORTGAGE	Verified	1	0.900115	0
216.17	20640.0	19.24	0.0	695.0	699.0	0.0	...	0.0	45.00	60.32	D	D1	OWN	Source Verified	1	0.897879	0
490.62	44250.0	13.34	1.0	660.0	664.0	1.0	...	0.0	0.00	489.85	F	F4	MORTGAGE	Source Verified	1	0.897817	0

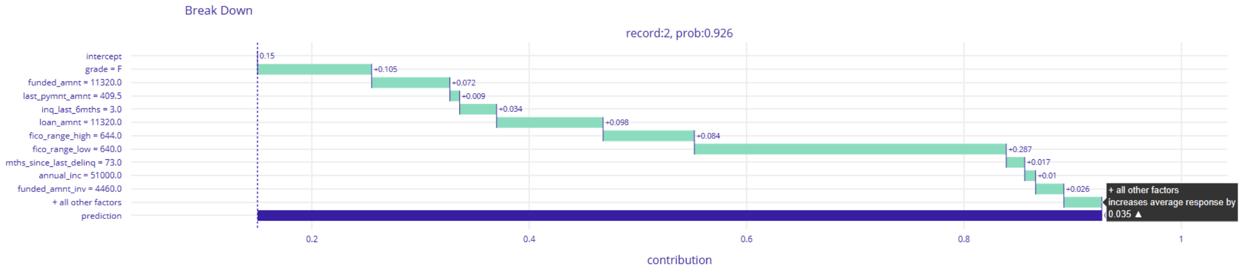
While this isn't a comprehensive table of the top 10 incorrectly predicted defaulted loans that did not default, we notice similarities in variables such as FICO range, home ownership, and loan grade/subgrade. The graphs below will further analyze select records, highlighting specific factors contributing to the probability of default.

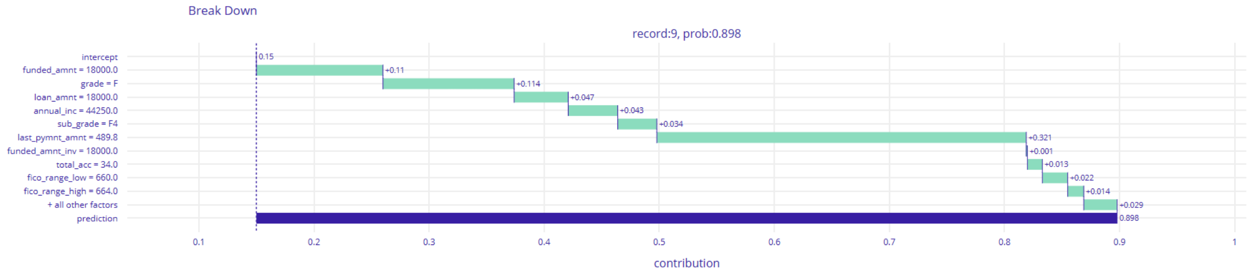


Reflecting on the cases of true positives (correct predictions), we observed that high late fees, low last payment amounts, and poor loan grades were significant contributors to defaulted loans. Here, the narrative remains largely similar, but the key difference is that these loans did not default. This prompts the question: what factors led to this misclassification?



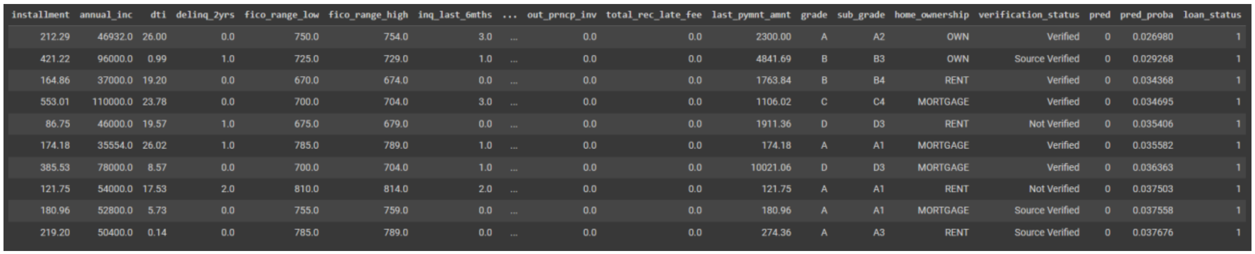
We encounter a similar scenario in this record, with significant contributions from various other factors. However, there are contradictions to our insights from variable importance. For instance, while mortgage home ownership was expected to increase the predicted probability of defaulting, here it indicates a decrease in the likelihood of defaulting. This discrepancy raises questions about the accuracy of our initial assumptions and highlights the need for further investigation.





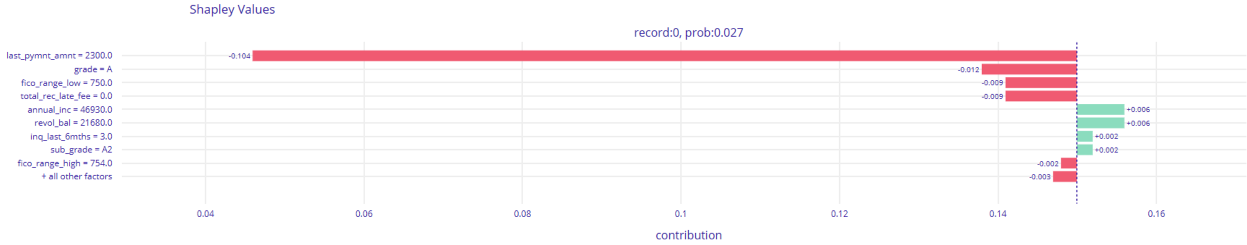
In reviewing the top 10 cases of false positives, we observe a sample breakdown revealing common drivers of predicted default probabilities. These include borrowers with poor loan grades, high loan and funded amounts, and significant contributions from various other factors. This pattern suggests potential model bias, where certain variables are disproportionately influencing predictions. It's possible that the model hasn't adequately learned about extreme cases where loans are non-defaulting despite indicators suggesting otherwise. Further investigation into these discrepancies is warranted to refine the model's accuracy.

False Negatives:

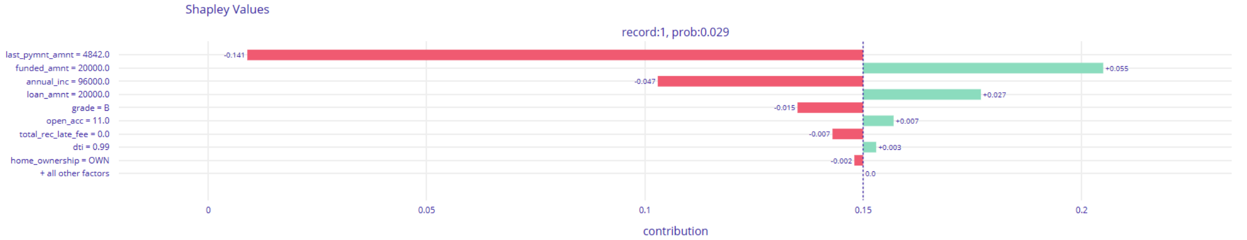


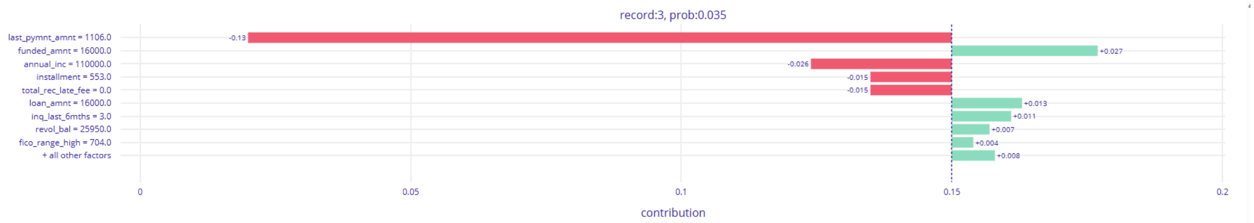
installment	annual_inc	dti	delinq_2yrs	fico_range_low	fico_range_high	inq_last_6mths	...	out_prncp_inv	total_rec_late_fee	last_pymnt_amnt	grade	sub_grade	home_ownership	verification_status	pred	pred_proba	loan_status
212.29	46932.0	26.00	0.0	750.0	754.0	3.0	...	0.0	0.0	2300.00	A	A2	OWN	Verified	0	0.026980	1
421.22	96000.0	0.99	1.0	725.0	729.0	1.0	...	0.0	0.0	4841.69	B	B3	OWN	Source Verified	0	0.029268	1
164.86	37000.0	19.20	0.0	670.0	674.0	0.0	...	0.0	0.0	1763.84	B	B4	RENT	Verified	0	0.034368	1
553.01	110000.0	23.78	0.0	700.0	704.0	3.0	...	0.0	0.0	1106.02	C	C4	MORTGAGE	Verified	0	0.034695	1
86.75	46000.0	19.57	1.0	675.0	679.0	0.0	...	0.0	0.0	1911.36	D	D3	RENT	Not Verified	0	0.035406	1
174.18	35554.0	26.02	1.0	785.0	789.0	1.0	...	0.0	0.0	174.18	A	A1	MORTGAGE	Verified	0	0.035582	1
385.53	78000.0	8.57	0.0	700.0	704.0	1.0	...	0.0	0.0	10021.06	D	D3	MORTGAGE	Verified	0	0.036363	1
121.75	54000.0	17.53	2.0	810.0	814.0	2.0	...	0.0	0.0	121.75	A	A1	RENT	Not Verified	0	0.037503	1
180.96	52800.0	5.73	0.0	755.0	759.0	0.0	...	0.0	0.0	180.96	A	A1	MORTGAGE	Source Verified	0	0.037558	1
219.20	50400.0	0.14	0.0	785.0	789.0	0.0	...	0.0	0.0	274.36	A	A3	RENT	Source Verified	0	0.037676	1

While this isn't a complete table of the top 10 incorrectly predicted non-defaulting loans that defaulted, we can observe similarities in values such as DTI, FICO range, total late fee, and loan grade/subgrade. The graphs below will delve into select records, highlighting specific factors that contributed to the probability of non-defaulting.

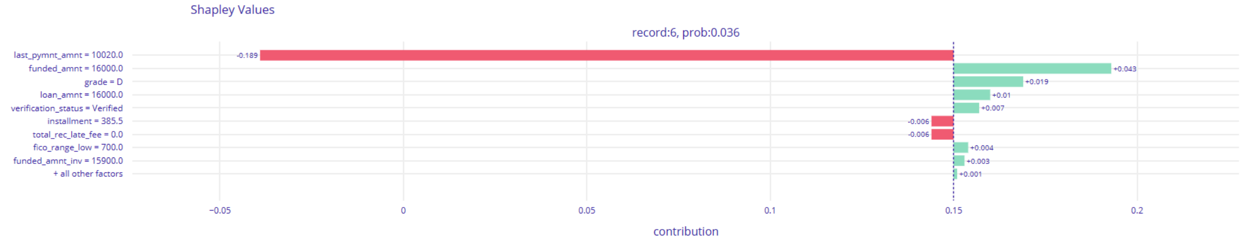


Our conventional understanding suggests that higher loan grades, such as A, decrease the chances of default, while high last payment amounts do not increase default likelihood. However, factors contributing to default include a loan subgrade of A2, a high revolving balance, and generally subpar annual income. These insights provide valuable context for predicting default probabilities.





This record underscores our model bias, where the last payment amount emerges as the primary driving factor for the predicted probability of a loan. Additionally, having a high annual income and a $0 total late fee outweigh factors that typically contribute to the predicted probability of a loan defaulting. This imbalance highlights the need to reassess our model's weighting of variables.



In examining the top 10 cases of false negatives, we observe a sample breakdown revealing common factors that have decreased predicted probabilities of defaulting. These include borrowers with excellent loan grades, low loan and funded amounts, and a $0 total late fee, all of which outweigh the factors that typically contribute to predicted default probabilities. Identifying where the model can falter, such as in false negatives and false positives, allows us to learn from these instances and make improvements to enhance the model's performance and accuracy in predicting loan defaults.