Project 3

Loan Defaults

Niklas Baldis

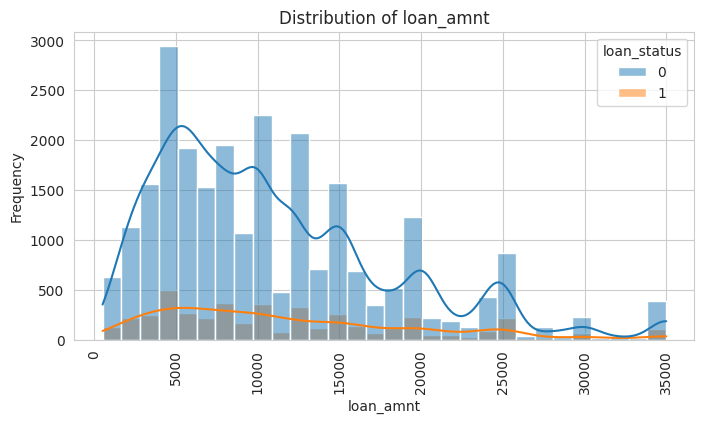
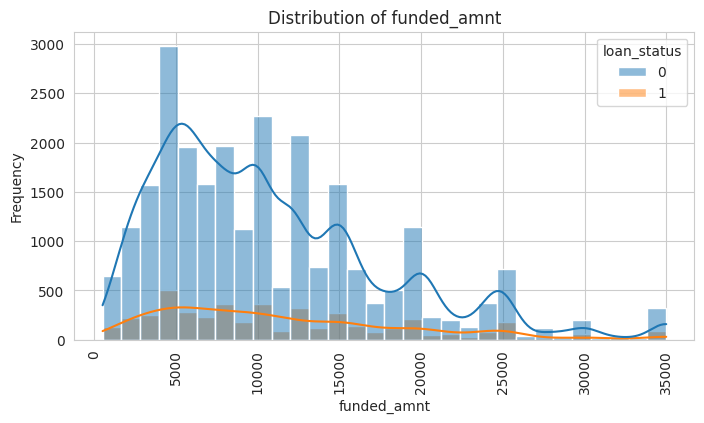
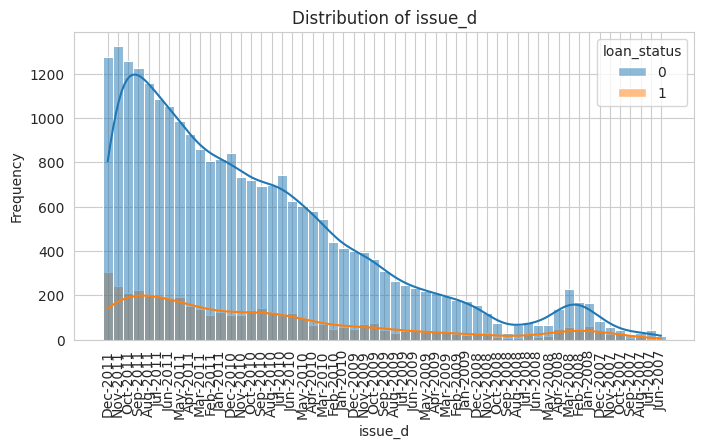
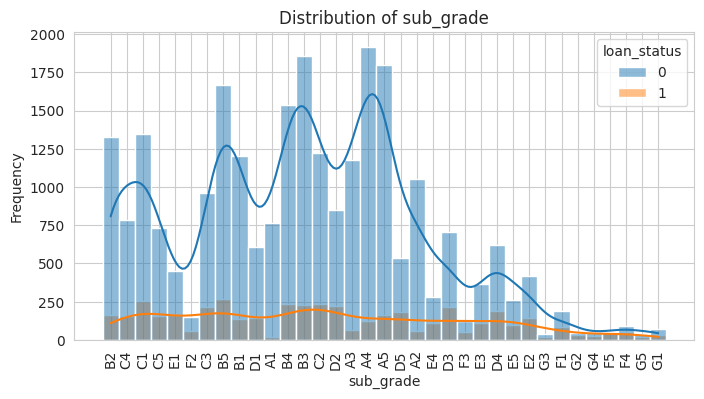
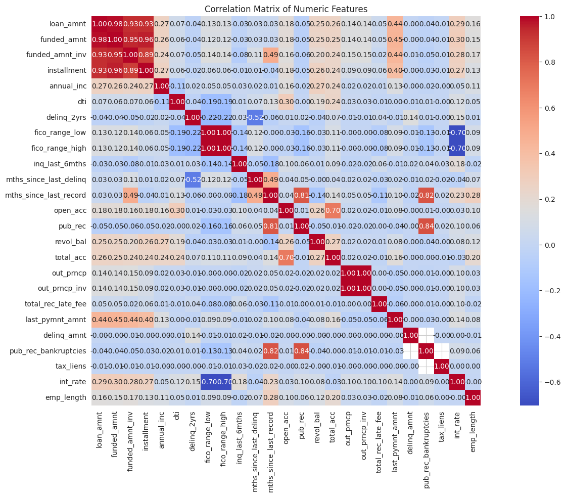
# Full Model Report

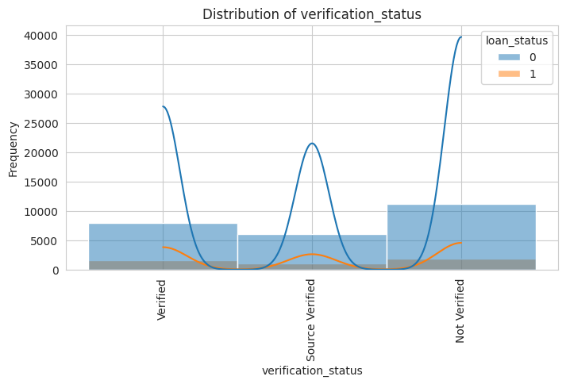
## Explorations Relative to Target

**Data Exploration and Preprocessing**

1. **Exploratory Data Analysis (EDA) & Feature Screening**: Conduct an initial analysis to understand the data's characteristics, including distribution of the target variable, missing values, and potential outliers.

The dataset includes both categorical and numerical columns. The "loan\_status" column is of particular interest, indicating whether an account has defaulted. This column trains the models to predict accounts that will default. During the exploratory analysis, it was found that $5000 is usually the most fraudulent loan amount for accounts, with a frequency of about 400 of the nearly 3000 total accounts given that loan amount. Approximately $5000 is also the most commonly invested amount for fraudulent accounts, with a frequency of about 350 of the nearly 1750 accounts with that investment amount. Many numerical features also exhibit a right skew, indicating that lower-volume and newer accounts are usually more fraudulent. Regarding correlation, Months since the last record and the number of derogatory public records are highly correlated. The correlation matrix also found that bankruptcies are highly correlated with months since the last record. The categorical features had much more random distributions of fraudulent accounts. However, a significant majority of the fraudulent accounts are found in the more recent months of the dataset, especially in 2011. This may be because loans were issued more frequently in 2010 and 2011. The Grades of “C” and “D,” with “C1,” “C2,” “B3,” and “B5” subgrades, were found to be the most fraudulent. Exploratory analysis also found that non-verified accounts are slightly more likely to be fraudulent.



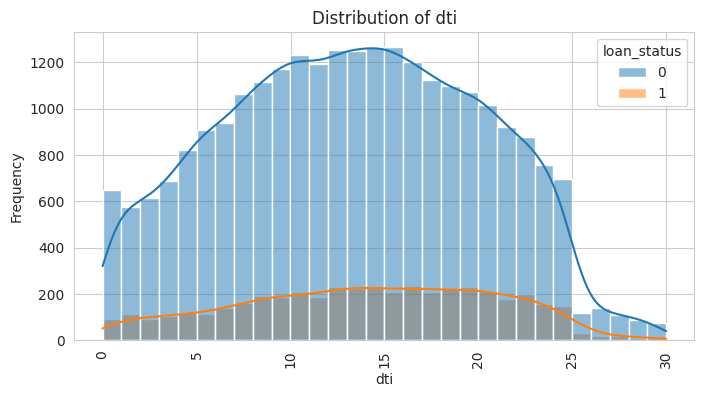
1. **Data Preprocessing**: Address missing values, encode categorical variables, and standardize numerical features to prepare the data for modeling.

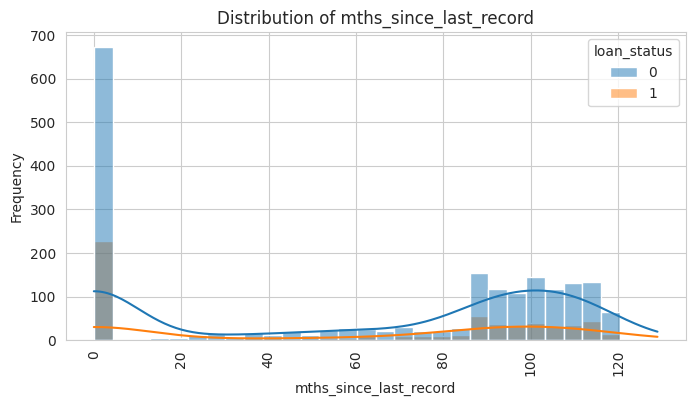
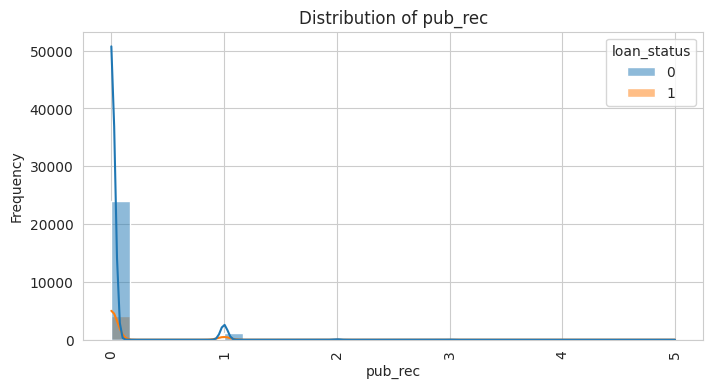
As shown in the bar plots, numerous potential outliers were observed on both ends of the variable relationships. Missing values in the “engineered” numerical columns were imputed using Pandas's fillna feature, using the column's median. Missing values in numerical features that were not engineered were imputed using the SimpleImputer feature of the Pipeline package. The same process was used for categorical columns, but they were instead filled with the constant “missing” for both “engineered” and “non-engineered” categorical features. The numerical data was scaled for the pipelines, and the categorical features were encoded using the OneHotEncoder feature of SkLearn’s preprocessing package.

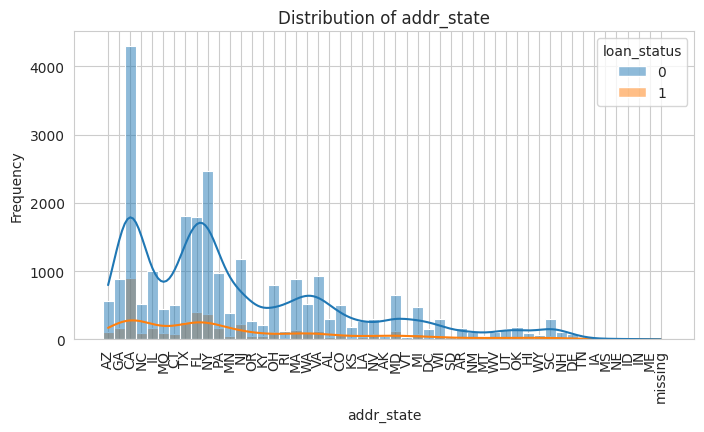
## Anomaly Detection

Some of the most significant anomalies compared to the rest of the dataset include the records that were loaned and funded between $30,000 and $35,000, those with Deb-to-Income ratios over 25%, accounts that recently received a derogatory public record (0 months), those that had one derogatory public record, and accounts that are in California. All of these outliers except the months since the last derogatory record and California accounts have significantly fewer accounts in the dataset than the rest of the features. These anomalistic features could substantially impact the imputation of data and bias the models to skew toward their outlier features.

A graph with blue and orange lines

Description automatically generated 



## Modeling Training

1. **Table of Performance**: Evaluate models using accuracy, AUC-ROC, precision, recall, and F1-score.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | Optimized Random  Forest | Optimized GBM | Neural Network | Stacking Classifier |
| ROC | 0.6297 | 0.8549 | **0.8992** | 0.8124 | 0.8971 |
| Precision | 0.6421 | 0.6520 | **0.7083** | 0.4396 | 0.6732 |
| Recall | 0.2872 | 0.1510 | 0.3666 | **0.3927** | 0.3905 |
| F1 | 0.3969 | 0.2452 | 0.4832 | 0.4149 | **0.4943** |

The F1 score is the “harmonic mean” of precision and recall. This means that it more fairly represents both precision and recall to evaluate a model’s viability more accurately.

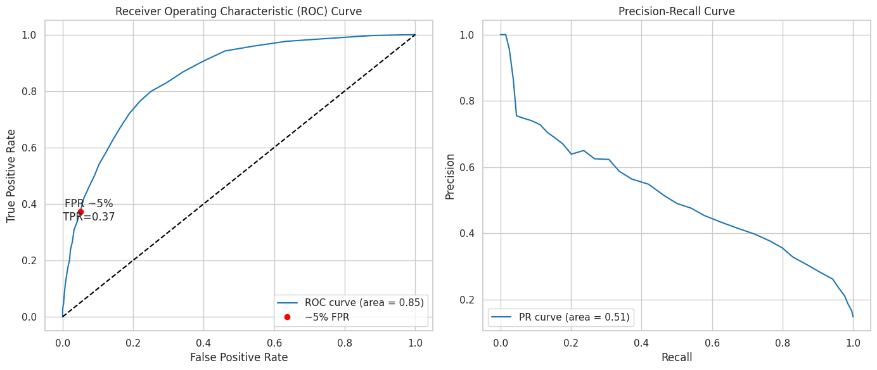
1. **ROC Charts with Score Threshold**

**Logistic**

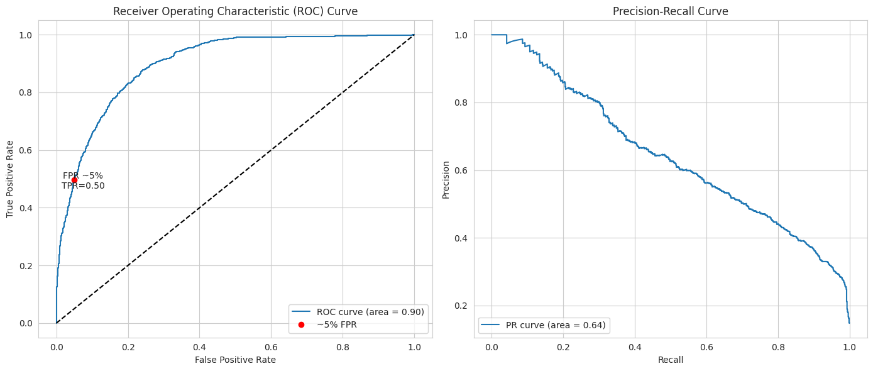
A graph of a graph of a graph

Description automatically generated with medium confidence

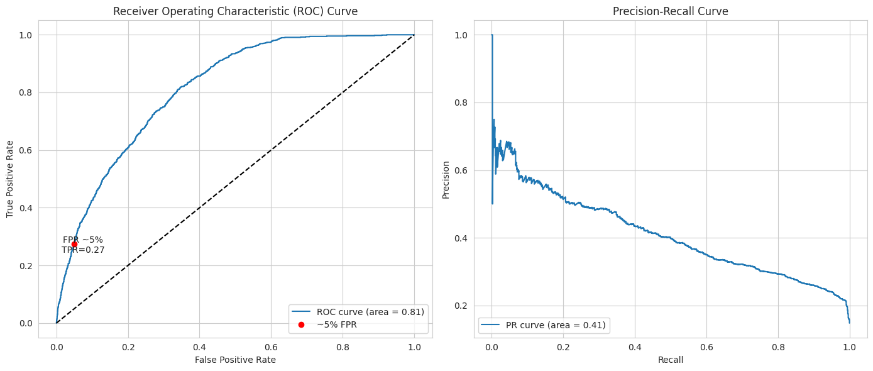
**Random Forest**



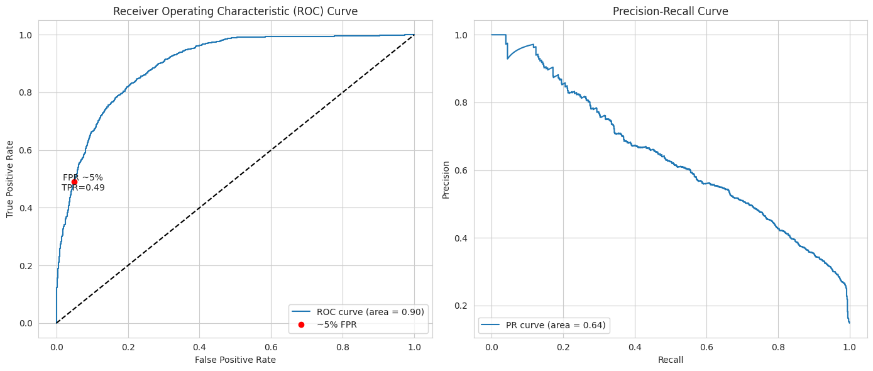
**GBM**



**Neural Network**



**Stacking Classifier**



Operating at a 5% false positive rate means that out of every 100 accounts flagged as legitimate, 5 defaulted accounts are falsely labeled as legitimate. The 5% threshold for logistic regression correlates with a 100% true positive rate. For Random Forest, the True positive rate goes to 37%, GBM calculates a TPR of 50%, and the Neural Network receives a TPR of 27%. Finally, the stacking classifier receives a TPR of 49%. Although there is a risk in allowing defaulted accounts to pass through the models, the model correctly identifies both defaulted and legitimate accounts at a 95% accuracy rate. However, the system still makes occasional mistakes, highlighting the need for fine-tuning to improve efficiency and customer experience. In this scenario, using a threshold of 0.389132 for predicted probability would capture about 50% of all defaulted accounts with a precision of 95% while incorrectly flagging 5% of legitimate accounts as defaulted.

1. **GBM Operating Table**

|  |  |  |
| --- | --- | --- |
| Target FPR (%) | Expected TPR | Threshold |
| 1 | 0.2588 | 0.641912 |
| 2 | 0.3292 | 0.544357 |
| 3 | 0.3814 | 0.485261 |
| 4 | 0.4393 | 0.438569 |
| 5 | 0.496027 | 0.389132 |
| 6 | 0.525539 | 0.362178 |
| 7 | 0.575482 | 0.326002 |
| 8 | 0.595914 | 0.301903 |
| 9 | 0.628831 | 0.282036 |
| 10 | 0.656073 | 0.262559 |

## How Did You Choose Hyperparameters?

Hyperparameters were chosen through the grid search method. Parameters of interest were selected to evaluate the best combination of parameters, and the parameters that produced the best model according to the ROC curve were chosen for the final “optimized” versions of the models. The Random Forest model hyperparameter optimization fit 54 different models, while the GBM fit 18.

## Global Explanations of Your Best Model

**Variable Importance Analysis**: Determine the most influential features in predicting defaulting accounts.

|  |  |
| --- | --- |
| Feature | Importance |
| last\_pymnt\_amnt | 0.408359 |
| total\_rec\_late\_fee | 0.153558 |
| int\_rate | 0.107309 |
| term\_60.0 | 0.050490 |
| Installment | 0.037000 |
| out\_prncp\_inv | 0.027973 |
| funded\_amnt\_inv | 0.026659 |
| term\_36.0 | 0.022851 |
| annual\_inc | 0.018435 |
| loan\_amnt | 0.015295 |

**GBM**

The model I chose was the GBM, as it had the highest ROC, and its precision and recall were near the top of the metrics for the models. ROC is one of the strongest metrics to evaluate models because it defines how well the model can distinguish between the two categories of default accounts and current accounts. This ability to accurately separate the groups leads to a stronger predictive model.

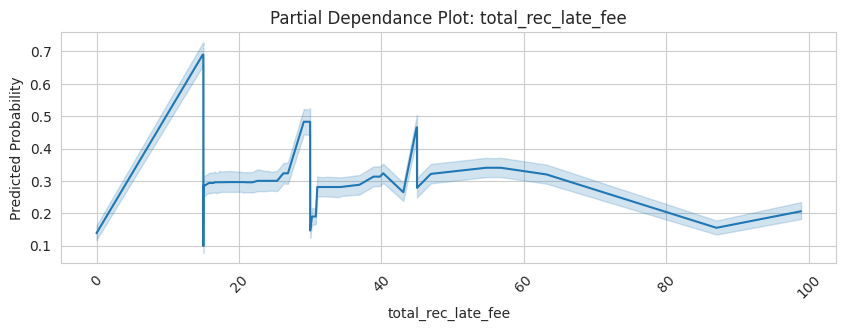
**Partial Dependency Plot of Top Variables**

**Last Payment Amount**

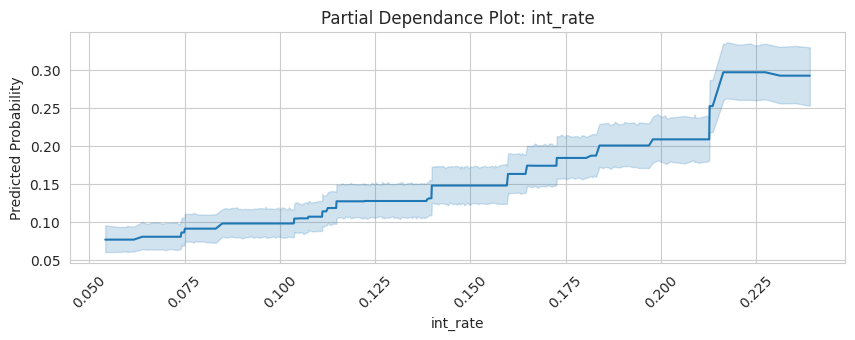
A graph with numbers and lines

Description automatically generated

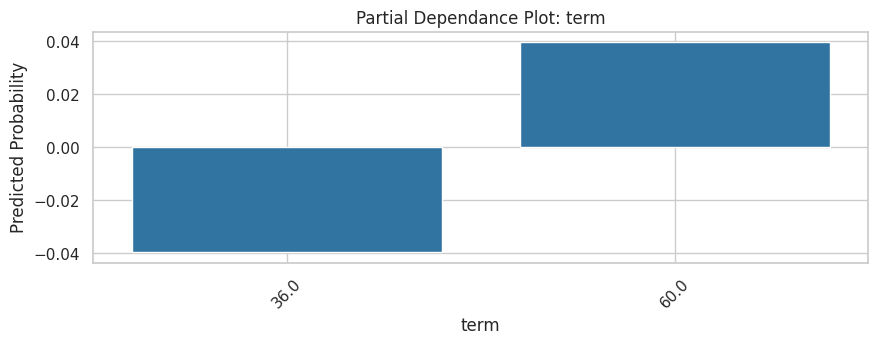
**Total Recorded Late Fee**



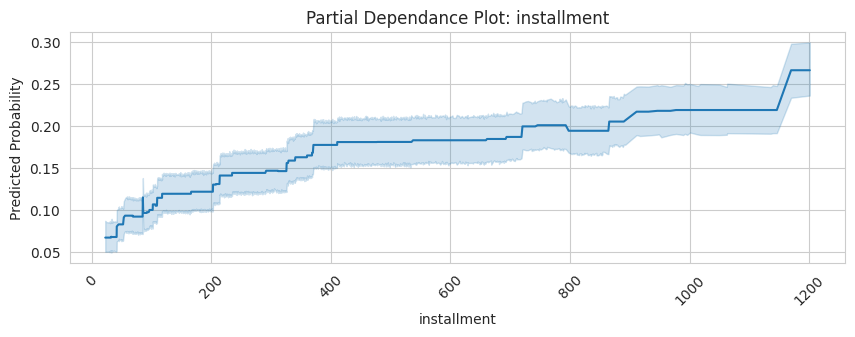
**Interest Rate**



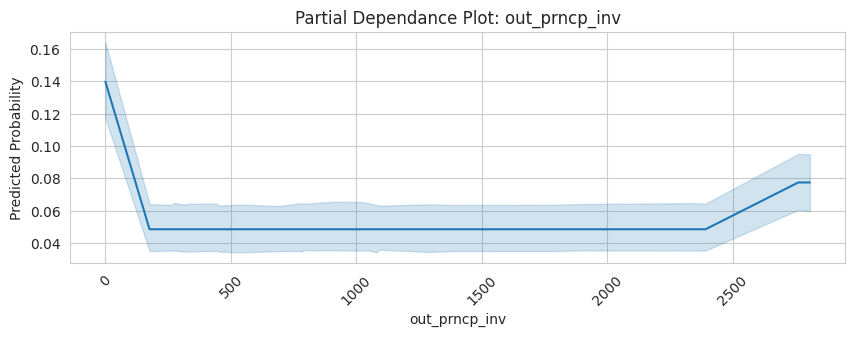
**60 Month Term**



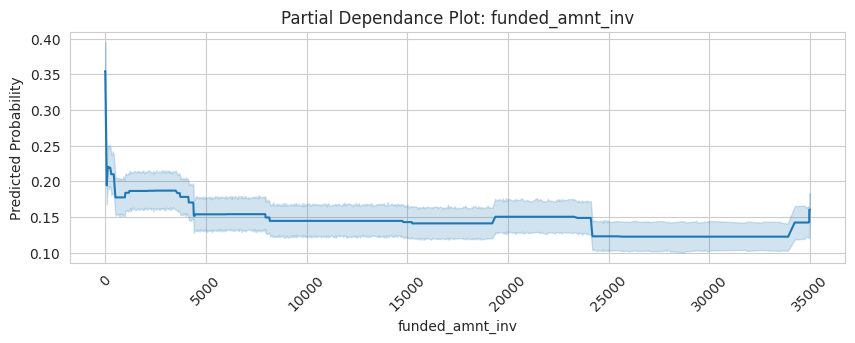
**Installment**



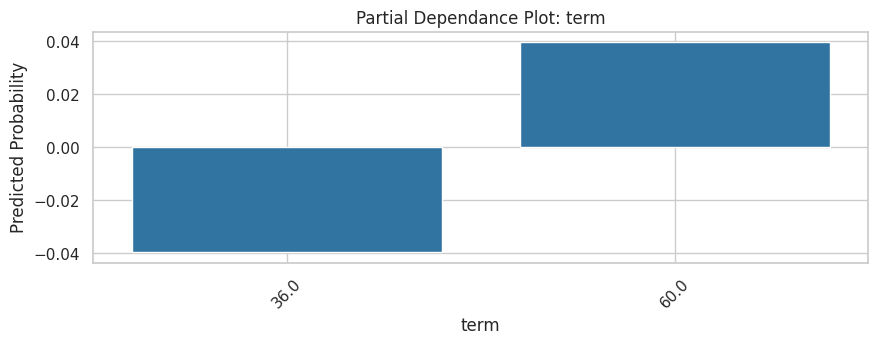
**Out\_Prncp\_Inv**



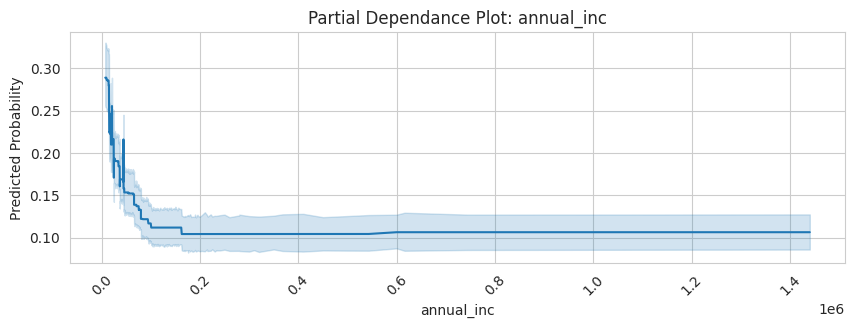
**Funded Amount Invested**



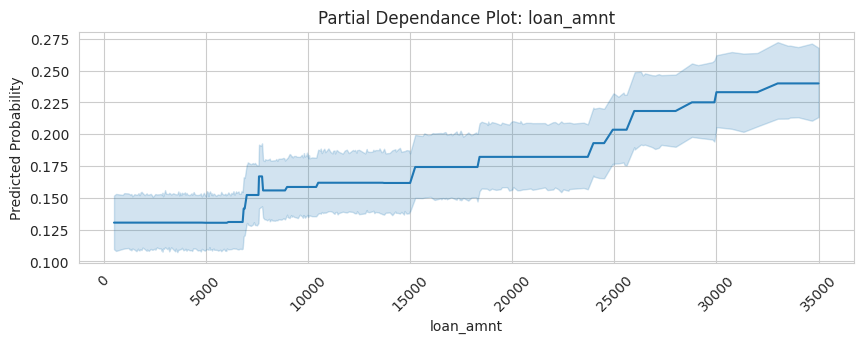
**36 Month Term**



**Annual Income**



**Loan Amount**



These variables are the most important to the GBM model’s predictions of defaulted accounts, as evidenced by the partial dependence graphs. These graphs show these features' impact on the model's predictive capabilities. This is especially shown in the last partial dependence graph on the loan amount, where there is a probability of fraud of nearly 80% for accounts that were recently paid. The loan amount PD plot shows how the variable is of tenth importance to the model, and there is incredible variation within the predicted probabilities.

## Local Explanations of Your Best Model

**Local Explanations**

* 1. **True Positives**

A graph with numbers and lines

Description automatically generated with medium confidenceA close-up of a graph

Description automatically generatedA graph with numbers and lines

Description automatically generated with medium confidenceA graph with green lines and white text

Description automatically generatedA close-up of a graph

Description automatically generatedA green and white graph

Description automatically generatedA graph with green and white lines

Description automatically generatedA graph with numbers and lines

Description automatically generated with medium confidenceA graph with green lines

Description automatically generated with medium confidenceA graph with numbers and lines

Description automatically generated with medium confidence

* 1. **False Positives**

A graph with green and white lines

Description automatically generatedA graph with green and white bars

Description automatically generatedA close-up of a graph

Description automatically generatedA graph with green and white bars

Description automatically generated with medium confidenceA bar graph with green and blue bars

Description automatically generatedA green and white bar graph

Description automatically generatedA bar graph with green and white bars

Description automatically generatedA graph with green and white bars

Description automatically generatedA green and white bar graph

Description automatically generatedA graph with numbers and lines

Description automatically generated with medium confidence

* 1. **False Negatives**

A graph with green and white lines

Description automatically generatedA green and white graph

Description automatically generatedA graph with green and white bars

Description automatically generatedA graph with green and white lines

Description automatically generatedA graph with multiple colored bars

Description automatically generated with medium confidenceA graph with green and white lines

Description automatically generatedA graph with green and white bars

Description automatically generatedA graph with multiple colored bars

Description automatically generated with medium confidenceA graph with green and white lines

Description automatically generatedA bar graph with green and white text

Description automatically generated