Source Code

CREDIT RISK SCORING

Problems statement

Based on data from 466,285 customers of a loan company, **50,968** customers have **defaulted on payments**

SOLUTION

Goals

Develop a machine learning credit risk scoring model to enhance the accuracy of lending decisions, reducing the likelihood of approving high-risk applicants while identifying creditworthy individuals

Objective

Create and train a machine learning model capable of predicting the creditworthiness of applicants, based on historical data and relevant features

Metrics

Measure the overall accuracy of the credit risk scoring model in correctly classifying applicants as high or low credit risk

Data Information

This data contains customer information from a loan company for the period 2007-2014

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	466285 non-null	int64
1	id	466285 non-null	int64
2	member_id	466285 non-null	int64
3	loan_amnt	466285 non-null	int64
4	funded_amnt	466285 non-null	int64
5	funded_amnt_inv	466285 non-null	float64
6	term	466285 non-null	object
7	int_rate	466285 non-null	float64
8	installment	466285 non-null	float64
9	grade	466285 non-null	object
10	sub_grade	466285 non-null	object
11	emp_title	438697 non-null	object
12	emp_length	445277 non-null	object
13	home_ownership	466285 non-null	object
14	annual_inc	466281 non-null	float64
15	verification_status	466285 non-null	object
16	issue_d	466285 non-null	object
17	loan_status	466285 non-null	object
18	pymnt_plan	466285 non-null	object
19	url	466285 non-null	object
20	desc	125983 non-null	object
21	purpose	466285 non-null	object
22	title	466265 non-null	object
23	zip_code	466285 non-null	object
24	addr_state	466285 non-null	object
25	dti	466285 non-null	float64
26	delinq_2yrs	466256 non-null	float64
27	earliest_cr_line	466256 non-null	object
28	inq_last_6mths	466256 non-null	float64
29	mths_since_last_delinq	215934 non-null	float64
30	mths_since_last_record	62638 non-null	float64
31	open_acc	466256 non-null	float64
32	pub_rec	466256 non-null	float64
33	revol_bal	466285 non-null	int64
34	revol_util	465945 non-null	float64
35	total_acc	466256 non-null	float64
36	initial_list_status	466285 non-null	object
37	out_prncp	466285 non-null	float64
38	out_prncp_inv	466285 non-null	float64
39	total_pymnt	466285 non-null	float64
40	total_pymnt_inv	466285 non-null	float64
41	total_rec_prncp	466285 non-null	float64
42	total_rec_int	466285 non-null	float64
43	total_rec_late_fee	466285 non-null	float64
44	recoveries	466285 non-null	float64
45	collection_recovery_fee	466285 non-null	float64
46	last_pymnt_d	465909 non-null	object
47	last_pymnt_amnt	466285 non-null	float64
48	next_pymnt_d	239071 non-null	object
49	last_credit_pull_d	466243 non-null	object
50	collections_12_mths_ex_med	466140 non-null	float64
51	mths_since_last_major_derog	98974 non-null	float64
52	policy_code	466285 non-null	int64
53	application_type	466285 non-null	object
54	annual_inc_joint	0 non-null	float64
55	dti_joint	0 non-null	float64

Data Information

- Columns mths_since_last_major_derog, mths_since_last_record, mths_since_last_delinq, and next_pymnt_d have a significant number of missing values.
- Columns emp_title and desc contain free-text data.
- Columns title, zip_code, sub_grade, and addr_state exhibit high cardinality with a large number of unique values.
- Columns emp_length can be impute with unemployee
- Columns total_rev_hi_lim, tot_cur_bal, tot_coll_amt, acc_now_delinq, collections_12_mths_ex_med, total_acc, revol_util, pub_rec, open_acc, inq_last_6mths, delinq_2yrs, annual_inc can be impute with the mean value

Define Target

It can be observed that the variable loan_status has several values:

- Current, which means payments are up-to-date
- Charged Off, which signifies defaulted payments that have been written off
- Late, indicating delayed payments; In Grace Period, indicating a grace period
- Fully Paid, meaning payments have been completed
- Default, which denotes defaulted payments

Secure Credit

- Current
- Fully paid
- Does not meet the credit policy. Status:Fully Paid

Risk Credit

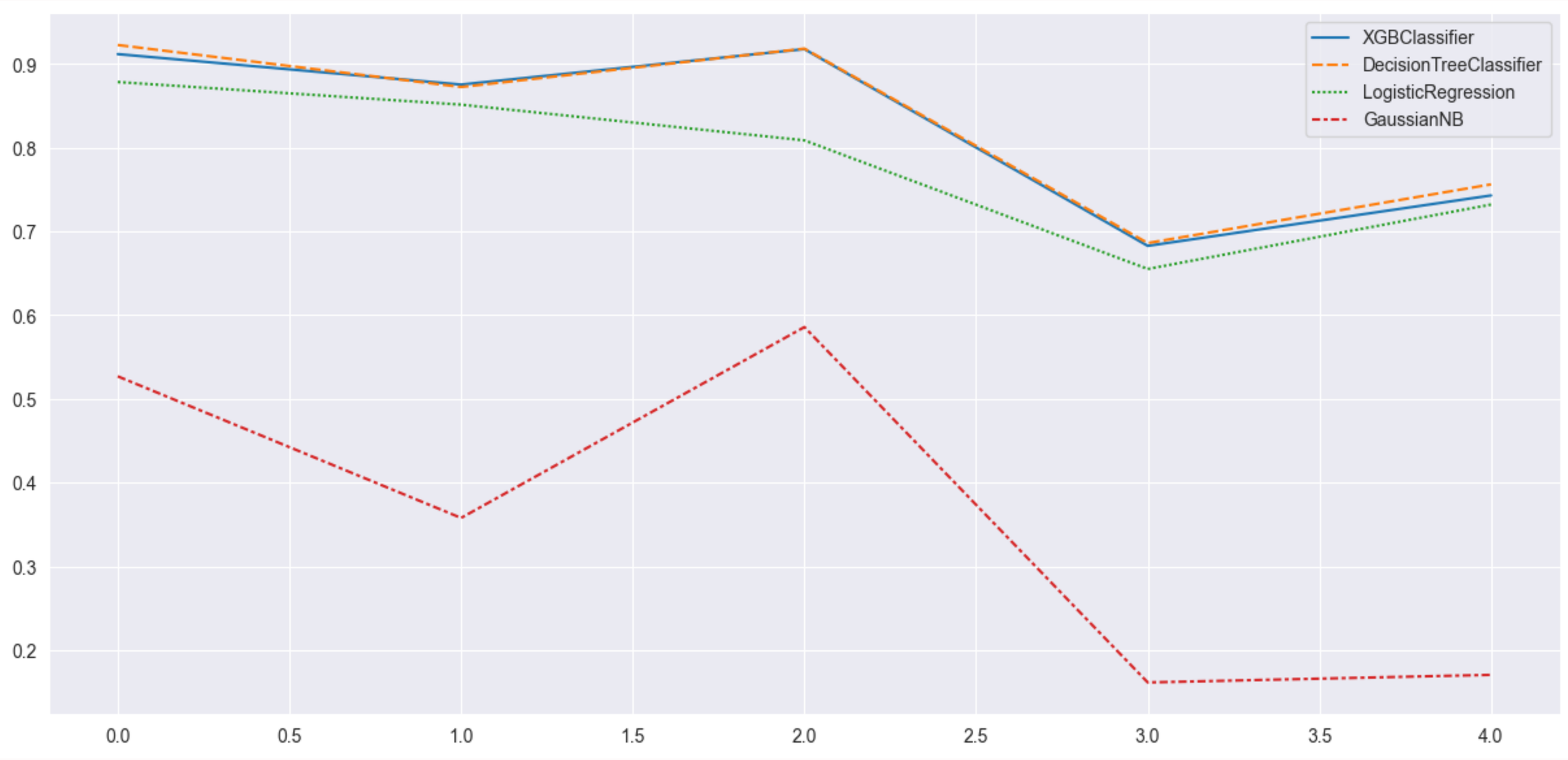
- Charged Off
- Default
- Does not meet the credit policy. status: charged off

Modelling

conduct training on the entire dataset using cross-validation. This is done to predict the outcome of the best model **recall** that will be utilized

	Trial 1	*Trial 2*	*Trial 3*	*Trial 4*	*Trial 5*
XGBoost Classifier	0.91	0.87	0.91	0.68	0.74
Decision Tree Classifier	0.92	0.87	0.91	0.68	0.75
Logistic Regression	0.87	0.85	0.80	0.65	0.73
Gaussan Naive Bayes	0.52	0.35	0.58	0.16	0.17

In this case, the **Decision Tree** and **XGBoost** model exhibits the highest average recall value among all other models, signifying its good performance in correctly identifying loan defaults.



From the results, it can be seen that the **Decision Tree** consistently has the highest performance in each trial, increasing the likelihood of using the **Decision Tree Classifier**

Hyperparameter Tuning

Before Hyperparameter tuning

Classification Report Model Default:

	precision	recall	f1-score	support
train	0.99	0.98	0.98	73541
test	0.83	0.86	0.85	7173
accuracy			0.97	80714
macro avg	0.91	0.92	0.92	80714
weighted avg	0.97	0.97	0.97	80714

After Hyperparameter tuning

Classification Report Model Hyperparameter Tuning:

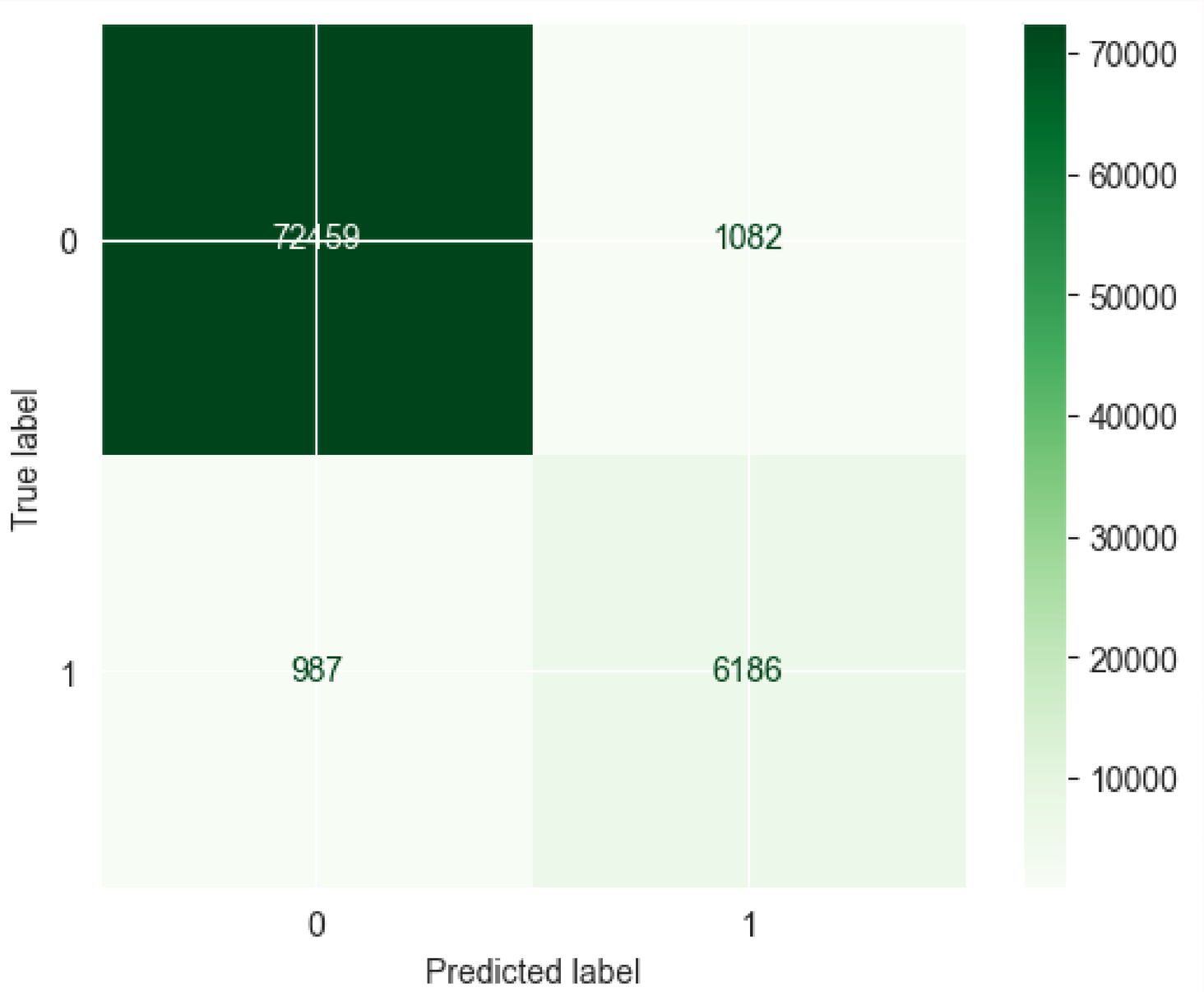
	precision	recall	f1-score	support
train	0.99	0.99	0.99	73541
test	0.85	0.86	0.86	7173
accuracy			0.97	80714
macro avg	0.92	0.92	0.92	80714
weighted avg	0.97	0.97	0.97	80714

Hyperparameter tuning, is the process of finding the best set of hyperparameters for a machine learning model. Hyperparameters are configuration settings for a machine learning algorithm that are not learned from the data but are set prior to training the model

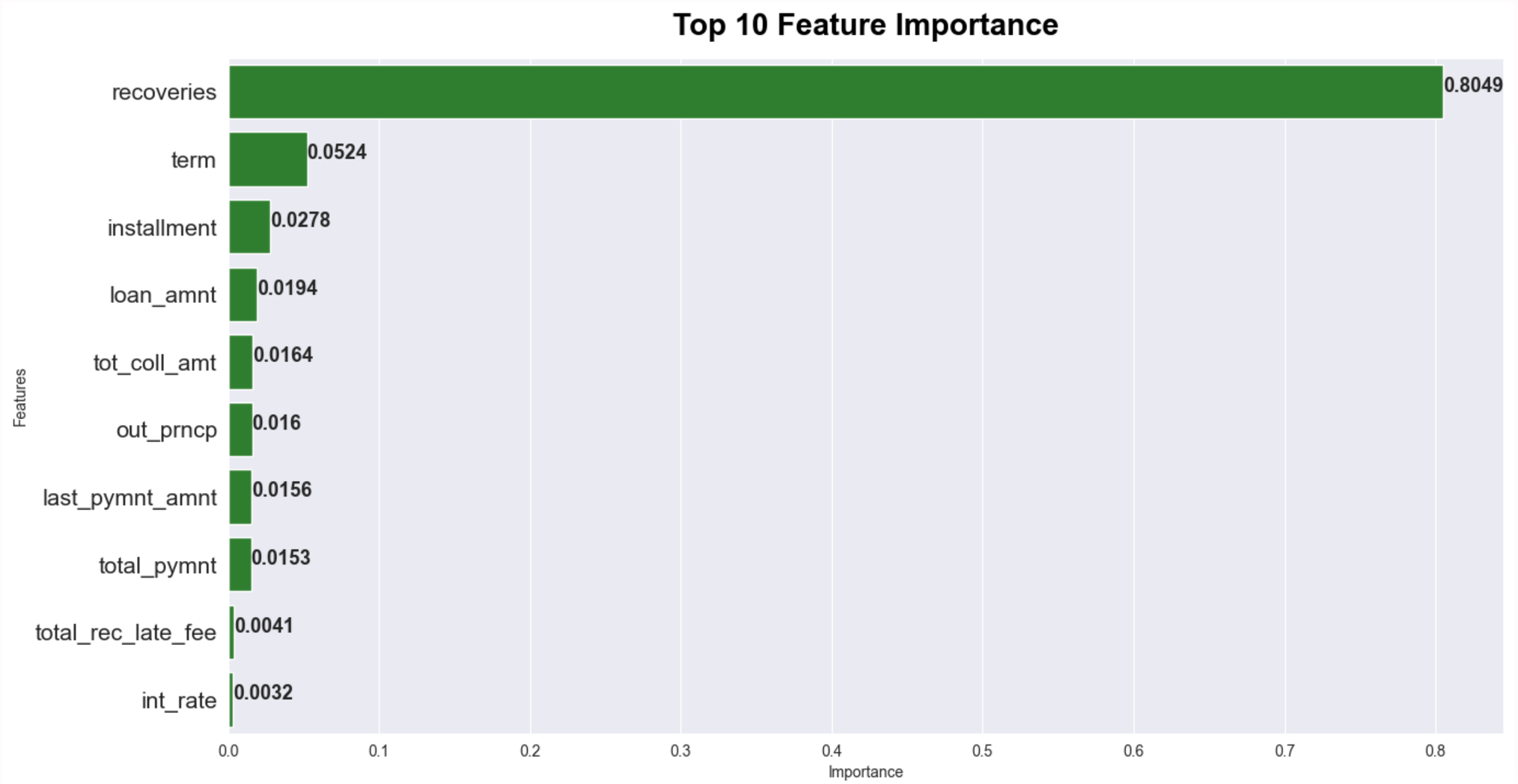
In the Classification Report results, it can be observed that after applying hyperparameter tuning, some of the performance metrics have **improved**, although not significantly.

Result

In this modeling outcome, there are **986 customers** classified as false positives with a loan default status. This result represents an improvement when compared to the pre-modeling data, which had **51,968 customers** identified as such.



Feature Importance



The columns that significantly influence credit risk scoring are recoveries (Indicates if a payment plan has been put in place for the loan), installment (The monthly payment owed by the borrower if the loan originates), and total payment (Last total payment amount received). These columns play a pivotal role in assessing the overall creditworthiness of applicants and have a substantial impact on the scoring process

ROC - AUC Curve

The constructed model yields a performance with an **AUC (Area Under the Curve) of 0.92**. In the realm of credit risk modeling, an AUC above 0.7 is generally considered a strong indicator of performance excellence.

This high AUC underscores the model's effectiveness in distinguishing between creditworthy and non-creditworthy applicants.