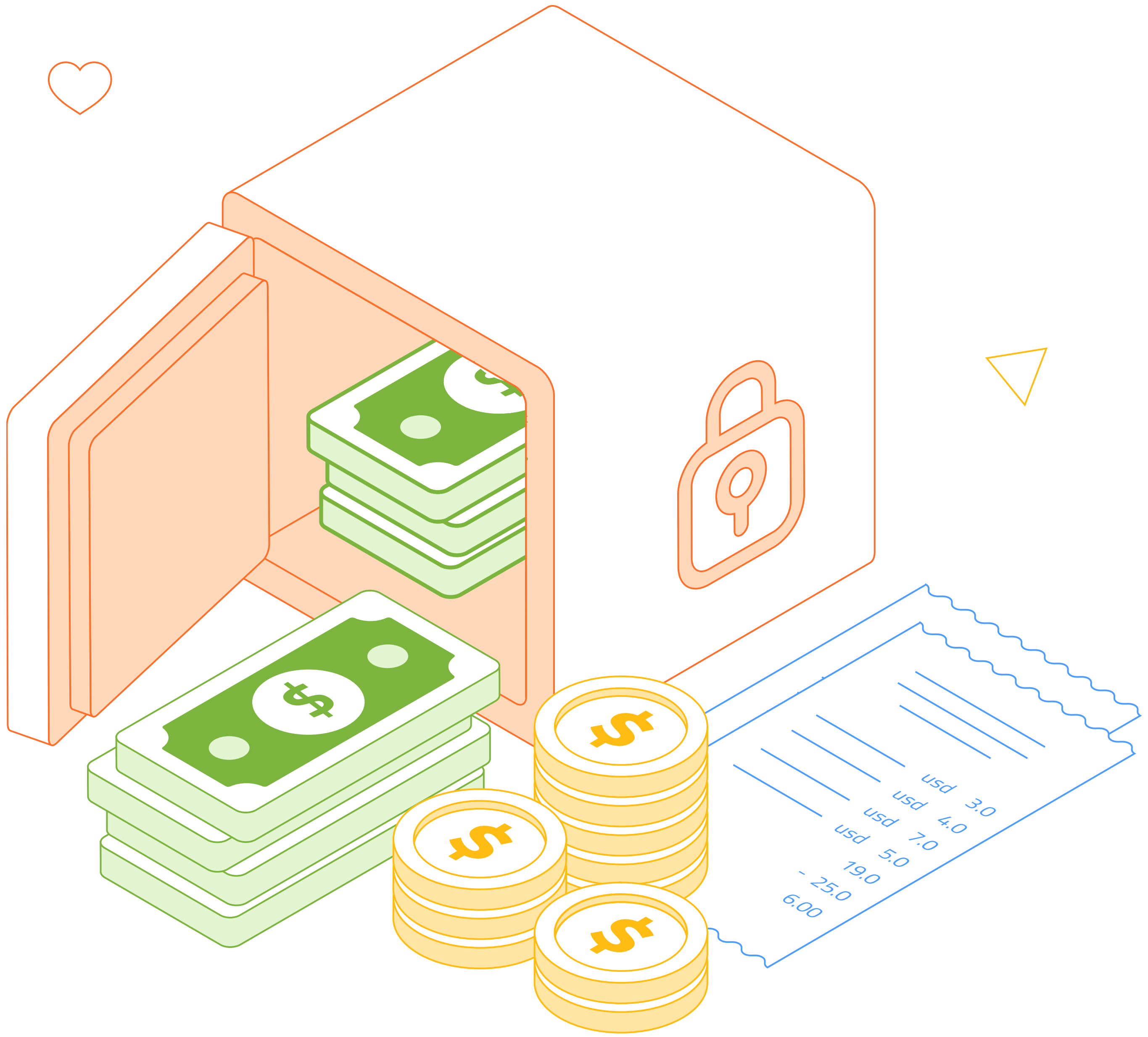


Predicting **Loan Default Risk** for Small Businesses

MEET THE TEAM!

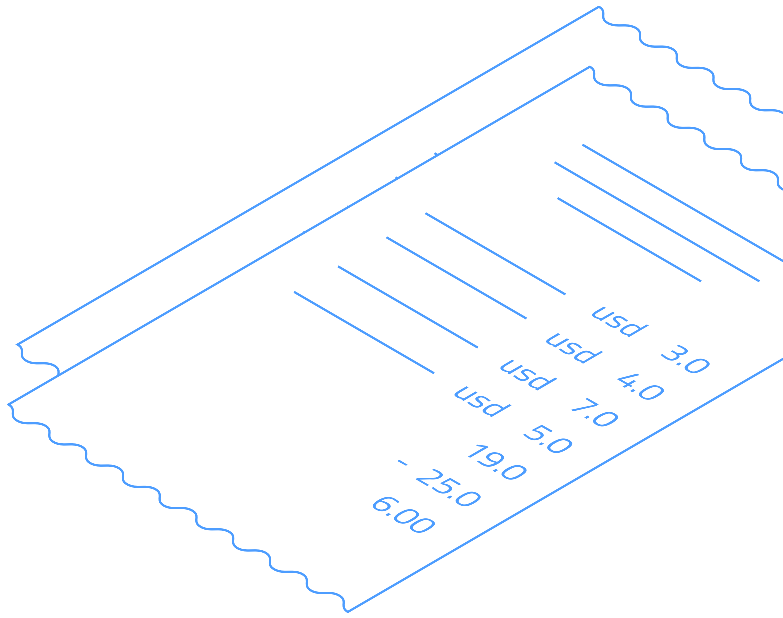
Jonathan Dang

Stella Lim

Andy Nguyen

Kaitlin Yen

TABLE OF CONTENTS



Data Cleaning

01

SBA Dataset

LoanNr_ChkI	Name	City	State	Zip	Bank	BankState	NAICS	ApprovalDate	ApprovalFY	Term	NoEmp	NewExist	CreateJob	RetainedJob	FranchiseCo	UrbanRural	RevLineCr	LowDoc	ChgOffDate	Disbursemer	Disbursemer	BalanceGros	MIS_Status	ChgOffPrinG	GrAppv	SBA_Appv
1E+09	ABC HOBBY	EVANSVILLE	IN	47711	FIFTH THIRD	OH	451120	28-Feb-97	1997	84	4	2	0	0	1	0	N	Y		28-Feb-99	#########	$0.00	P I F	$0.00	#########	#########
1E+09	LANDMARK E	NEW PARIS	IN	46526	1ST SOURCE	IN	722410	28-Feb-97	1997	60	2	2	0	0	1	0	N	Y		31-May-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	WHITLOCK D	BLOOMINGT	IN	47401	GRANT COU	IN	621210	28-Feb-97	1997	180	7	1	0	0	1	0	N	N		31-Dec-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	BIG BUCKS F	BROKEN AR	OK	74012	1ST NATL BK	OK	0	28-Feb-97	1997	60	2	1	0	0	1	0	N	Y		30-Jun-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	ANASTASIA C	ORLANDO	FL	32801	FLORIDA BU	FL	0	28-Feb-97	1997	240	14	1	7	7	1	0	N	N		14-May-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	B&T SCREW	PLAINVILLE	CT	6062	TD BANK, NA	DE	332721	28-Feb-97	1997	120	19	1	0	0	1	0	N	N		30-Jun-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	MIDDLE ATL	UNION	NJ	7083	WELLS FARG	SD	0	2-Jun-80	1980	45	45	2	0	0	0	0	N	N	24-Jun-91	22-Jul-80	#########	$0.00	CHGOFF	#########	#########	#########
1E+09	WEAVER PR	SUMMERFIEL	FL	34491	REGIONS BA	AL	811118	28-Feb-97	1997	84	1	2	0	0	1	0	N	Y		30-Jun-98	#########	$0.00	P I F	$0.00	#########	#########
1E+09	TURTLE BEA	PORT SAINT J	FL	32456	CENTENNIAL	FL	721310	28-Feb-97	1997	297	2	2	0	0	1	0	N	N		31-Jul-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	INTEXT BUIL	GLASTONBU	CT	6073	WEBSTER BA	CT	0	28-Feb-97	1997	84	3	2	0	0	1	0	N	Y		30-Apr-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	COMMERCIA	CHARLOTTE	NC	28256	SUNTRUST B	GA	811111	28-Feb-97	1997	84	1	2	0	0	1	0	N	Y		23-Feb-98	#########	$0.00	P I F	$0.00	#########	#########
1E+09	PROFESSION	CHICAGO	IL	60605	BANK OF AM	OR	235950	28-Feb-97	1997	60	24	1	0	0	1	0	N	N		30-Nov-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	CARVEL	APEX	NC	27502	STEARNS BK	MN	445299	7-Feb-06	2006	162	2	2	0	0	15100	1	N	N		31-Mar-06	#########	$0.00	P I F	$0.00	#########	#########
1E+09	ORCHARD C	SLATERSVILL	RI	2876	CITIZENS BA	RI	0	28-Feb-97	1997	120	2	2	0	0	1	0	N	N		31-May-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	EBC INVESTN	WINSTON-SA	NC	27106	NORTHWEST	NC	0	28-Feb-97	1997	240	1	1	30	0	1	0	N	N		17-Dec-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	ENVIRONME	OKLAHOMA	OK	73112	BANK OF AM	NC	421330	28-Feb-97	1997	12	5	2	0	0	1	0	N	N		30-Sep-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	ARK MAMAG	MIDLAND	TX	79701	WELLS FARG	TX	0	28-Feb-97	1997	60	5	1	0	0	1	0	N	Y		30-Jun-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	FAIRFAX CO	CENTREVILL	VA	20120	BANK OF AM	MD	0	28-Feb-97	1997	60	16	1	0	0	1	0	N	Y		31-Jul-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	FANTASTIC S	PLANO	TX	75093	NEWTEK SMA	NY	0	28-Feb-97	1997	84	12	2	0	0	19755	0	N	Y		31-May-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	SIR GOONY'S	KNOXVILLE	TN	37922	CITIZENS NA	TN	0	28-Feb-97	1997	120	4	2	0	0	1	0	N	Y		31-May-98	#########	$0.00	P I F	$0.00	#########	#########
1E+09	ECONOLOD	DUMAS	TX	79029	BUSINESS LO	SC	0	28-Feb-97	1997	300	12	2	0	0	1	0	N	N		30-Apr-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	YOUNG ACH	CORAL SPRI	FL	33065	BANESCO US	FL	624410	28-Feb-97	1997	87	2	1	0	0	1	0	N	N		31-Aug-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	NICOLES RE	JOHNSTON	RI	2919	BANK OF AM	RI	0	28-Feb-97	1997	114	6	1	0	0	1	0	N	N		31-Mar-98	#########	$0.00	P I F	$0.00	#########	#########
1E+09	TRIANGLE M	EULESS	TX	76040	FIRST BANK	TX	0	28-Feb-97	1997	144	90	1	0	0	1	0	N	N		30-Apr-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	SUBWAY	LITTLE ROCK	AR	72223	HOPE FCU	MS	722211	7-Feb-06	2006	126	7	1	0	0	1	1	N	N		30-Apr-06	#########	$0.00	P I F	$0.00	#########	#########
1E+09	DEE'S CORN	SAINT PETER	MN	56082	WELLS FARG	MN	451110	28-Feb-97	1997	60	2	1	0	0	1	0	N	N		30-Apr-97	#########	$0.00	P I F	$0.00	#########	#########
1E+09	C & S TRANS	INDEPENDEN	MO	64055	BANK OF AM	NC	0	28-Feb-97	1997	60	2	2	0	0	1	0	0	N		31-May-98	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	HUNTERSBR	MARSHFIELD	MA	2050	ROCKLAND T	MA	0	28-Feb-97	1997	240	3	1	0	0	1	0	N	Y		31-Jul-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	WEYLAND CO	CAMARILLO	CA	93010	WELLS FARG	SD	611110	7-Feb-06	2006	83	18	2	5	23	1	1	Y	N		28-Feb-06	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	SCROOGE'S	ANDERSON	SC	29621	CERTIFIED D	SC	445310	28-Feb-97	1997	240	1	2	5	0	1	0	N	N		14-Jan-98	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	CHICAGO BR	MIAMI	FL	33186	CITIBANK, N.	FL	238140	7-Feb-06	2006	84	4	1	0	4	1	1	Y	N		28-Feb-06	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	AUDELIA FAN	DALLAS	TX	75243	THE FROST N	TX	621210	28-Feb-97	1997	102	12	1	0	0	1	0	N	N		31-Jul-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	RZI, INC.	NEW ORLEA	LA	70130	BUSINESS RE	LA	532490	7-Feb-06	2006	60	3	1	0	0	1	1	N	N		31-May-06	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	PPP COMMU	WASHINGTO	IA	52353	WASHINGTO	IA	454210	28-Feb-97	1997	84	2	2	0	0	1	0	N	N		31-Oct-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	HUTMACHER	LEANDER	TX	78641	WELLS FARG	SD	541611	7-Feb-06	2006	80	2	1	4	6	1	2	Y	N		31-May-06	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	PRESTIGE LIN	ROANOKE	VA	24015	FIRST COMM	VA	0	28-Feb-97	1997	84	2	1	0	0	1	0	N	Y		14-Mar-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	PAUL E. & JU	KINSMAN	OH	44428	CORTLAND S	OH	0	28-Feb-97	1997	137	2	1	0	0	1	0	N	Y	18-Apr-02	30-Jun-97	#########	$0.00	CHGOFF	#########	#########	#########
1.001E+09	VILLAGE RES	NORTH EAST	MA	2356	HOME LOAN	RI	0	28-Feb-97	1997	84	9	1	0	0	1	0	N	Y		31-May-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	CORBIN CRE	SPRINGFIELD	TN	37172	BBCN BANK	CA	453110	7-Feb-06	2006	84	4	1	1	4	1	1	0	N		28-Feb-06	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	JFJ PROCESS	LEWISBURG	TN	37091	FIRST FARME	TN	311611	28-Feb-97	1997	180	7	1	0	0	1	0	0	N		30-Apr-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	M.A.S. TRUC	SPRINGFIELD	IL	62702	PNC BANK, N	IL	0	28-Feb-97	1997	84	3	1	0	0	1	0	0	N		31-May-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	OLD LOUISV	LOUISVILLE	KY	40208	PNC BANK, N	KY	0	28-Feb-97	1997	126	5	1	0	0	1	0	0	N		31-May-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	IRON HORSE	LELAND	MS	38756	STATE BANK	MS	332996	28-Feb-97	1997	120	2	2	0	0	1	0	N	Y		30-Jun-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	LARRY SCHO	EDINBURGH	IN	46124	JPMORGAN C	IN	0	11-Jun-80	1980	120	16	2	0	0	0	0	Y	N	4-Oct-89	31-Jul-80	#########	$0.00	CHGOFF	#########	#########	#########
1.001E+09	Sun Service	Newburgh	NY	12550	WELLS FARG	SD	0	4-Oct-96	1997	84	3	1	0	0	1	0	0	N		31-Jul-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	Dover Quality	Dover (censu	MA	2030	BANK OF AM	MA	0	25-Mar-97	1997	12	20	1	0	0	1	0	0	N		30-Sep-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	SNADER EXC	SMITHVILLE	OH	44677	FIRST NATIO	OH	235930	28-Feb-97	1997	84	1	1	0	0	1	0	N	Y		31-Mar-97	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	RAYMIES GR	Chicago	IL	60628	WELLS FARG	SD	0	25-Mar-97	1997	84	4	1	0	0	1	0	0	N		31-Jul-99	#########	$0.00	P I F	$0.00	#########	#########
1.001E+09	ANYWHERE	Marina del R	CA	90292	WELLS FARG	SD	0	25-Mar-97	1997	84	6	1	0	0	1	0	Y	N		31-Oct-97	#########	$0.00	P I F	$0.00	#########	#########

Data Cleaning Process

01 Quality Check

Drop null values
Adjust data type
(integer vs. object)

02 Simplify Variables

NAICS code
States
NewExist

03 Convert to Binary

Change variables to be interpreted by the LRM

04 Remove & Rename

Remove inconsistent entries and unnecessary columns
Rename for readability

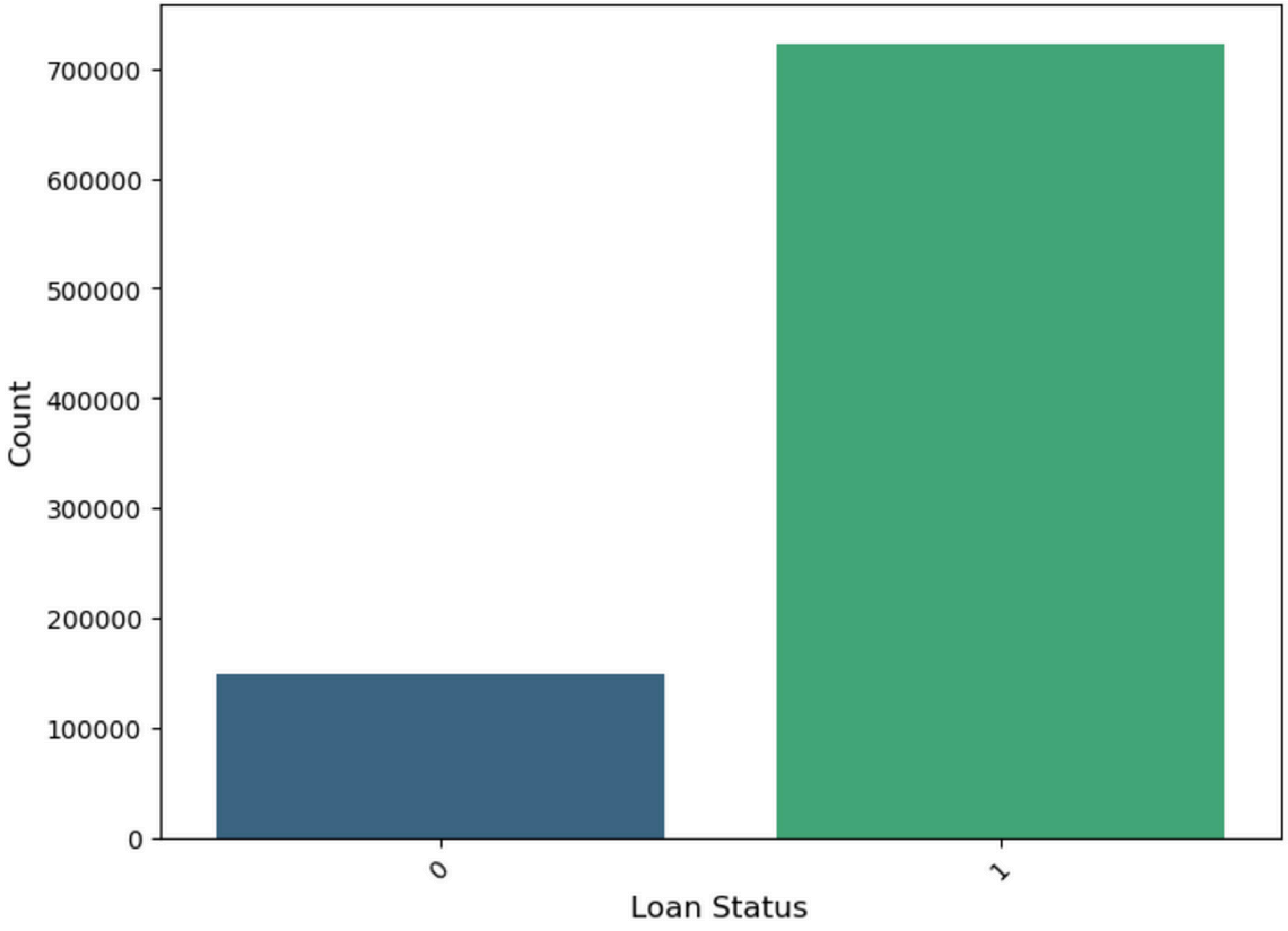
Final Dataset

NAICS_U.S._I	Loan_Term	Number_Of_I	Business_Typ	Urban_Rural_	Revolving_Lir	Low_Docume	Disbursemer	Loan_Status	Franchise_St	Month_Of_Ap	Region
RetailTrade	84	4	New	Unknown	0	1	60000	1	0	February	Midwest
Accommodat	60	2	New	Unknown	0	1	40000	1	0	February	Midwest
HealthCare_	180	7	Existing	Unknown	0	0	287000	1	0	February	Midwest
Unknown	60	2	Existing	Unknown	0	1	35000	1	0	February	South
Unknown	240	14	Existing	Unknown	0	0	229000	1	0	February	South
Manufacturir	120	19	Existing	Unknown	0	0	517000	1	0	February	Northeast
Unknown	45	45	New	Unknown	0	0	600000	0	0	June	Northeast
OtherService	84	1	New	Unknown	0	1	45000	1	0	February	South
Accommodat	297	2	New	Unknown	0	0	305000	1	0	February	South
Unknown	84	3	New	Unknown	0	1	70000	1	0	February	Northeast
OtherService	84	1	New	Unknown	0	1	70000	1	0	February	South
Constructior	60	24	Existing	Unknown	0	0	150000	1	0	February	Midwest
RetailTrade	162	2	New	Urban	0	0	253400	1	1	February	South
Unknown	120	2	New	Unknown	0	0	370000	1	0	February	Northeast
Unknown	240	1	Existing	Unknown	0	0	225000	1	0	February	South
WholesaleTr	12	5	New	Unknown	0	0	350000	1	0	February	South
Unknown	60	5	Existing	Unknown	0	1	70000	1	0	February	South
Unknown	60	16	Existing	Unknown	0	1	100000	1	0	February	South
Unknown	84	12	New	Unknown	0	1	57500	1	1	February	South
Unknown	120	4	New	Unknown	0	1	50000	1	0	February	South
Unknown	300	12	New	Unknown	0	0	615000	1	0	February	South
HealthCare_	87	2	Existing	Unknown	0	0	70000	1	0	February	South
Unknown	114	6	Existing	Unknown	0	0	75000	1	0	February	Northeast
Unknown	144	90	Existing	Unknown	0	0	1250000	1	0	February	South
Accommodat	126	7	Existing	Urban	0	0	137300	1	0	February	South
RetailTrade	60	2	Existing	Unknown	0	0	39500	1	0	February	Midwest
Unknown	60	2	New	Unknown	0	0	50000	1	0	February	Midwest
Unknown	240	3	Existing	Unknown	0	1	75000	1	0	February	Northeast
Educational	83	18	New	Urban	1	0	438541	1	0	February	West
RetailTrade	240	1	New	Unknown	0	0	291000	1	0	February	South
Constructior	84	4	Existing	Urban	1	0	51440	1	0	February	South
HealthCare_	102	12	Existing	Unknown	0	0	600000	1	0	February	South
RealEstate_F	60	3	Existing	Urban	0	0	50000	1	0	February	South
RetailTrade	84	2	New	Unknown	0	0	30000	1	0	February	Midwest
Professional_	80	2	Existing	Rural	1	0	63076	1	0	February	South
Unknown	84	2	Existing	Unknown	0	1	60000	1	0	February	South
Unknown	137	2	Existing	Unknown	0	1	47000	0	0	February	Midwest
Unknown	84	9	Existing	Unknown	0	1	70000	1	0	February	Northeast
RetailTrade	84	4	Existing	Urban	0	0	20000	1	0	February	South

Visualizations

02

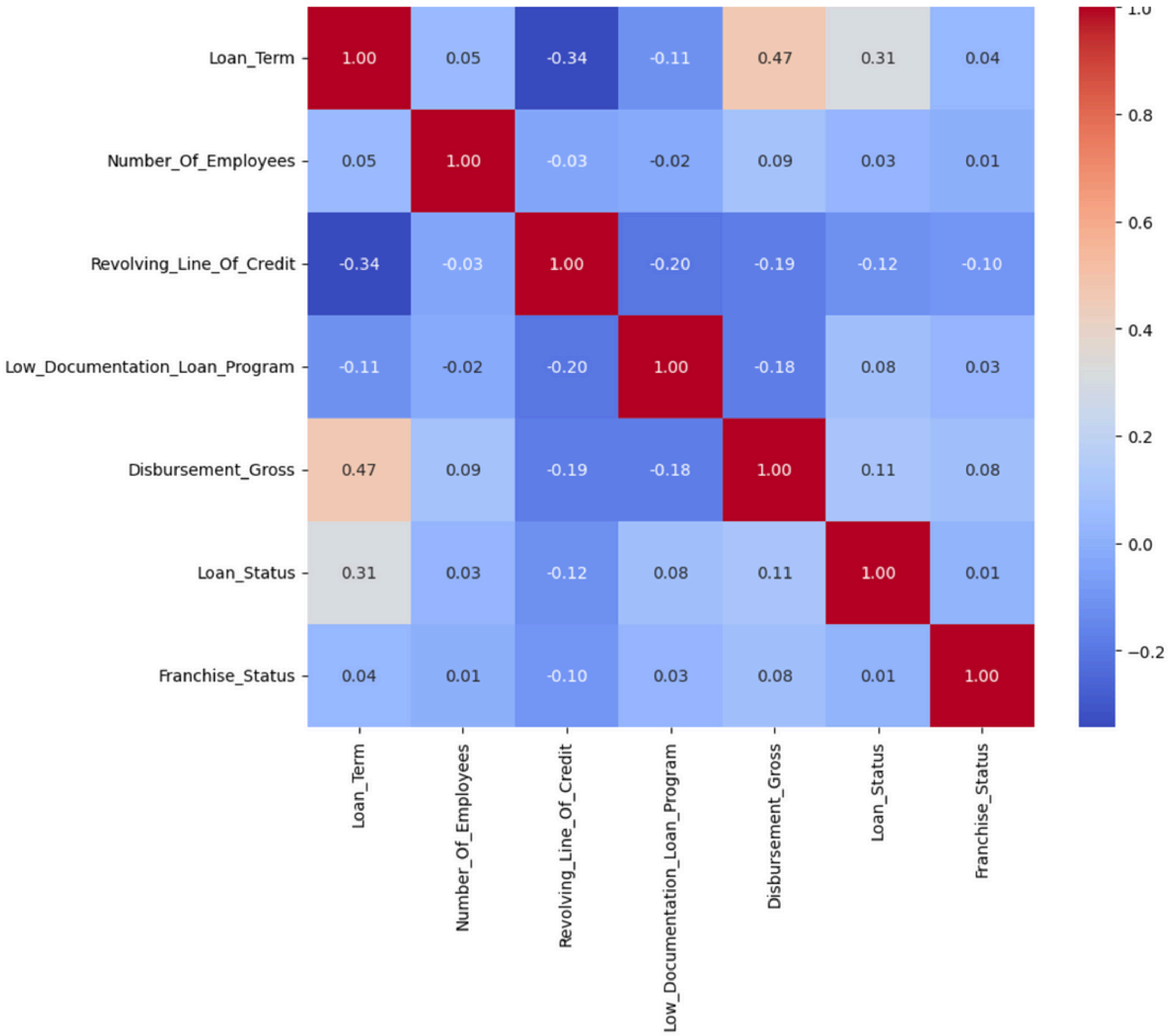
Loan Status Distribution



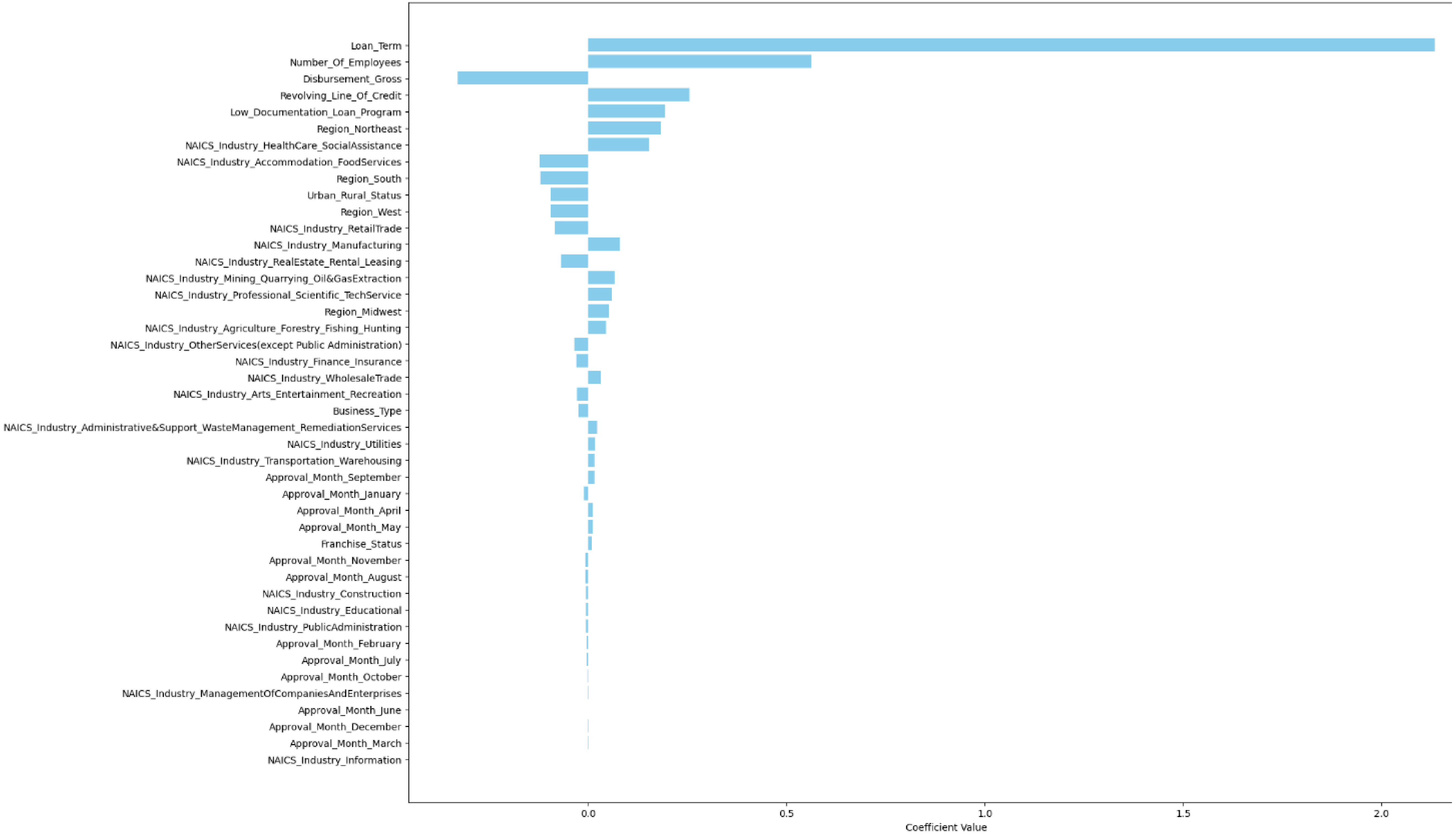
Correlation Heatmap

Takeaway

Moderate correlation between **loan term** & **amount disbursed**



Feature Importance

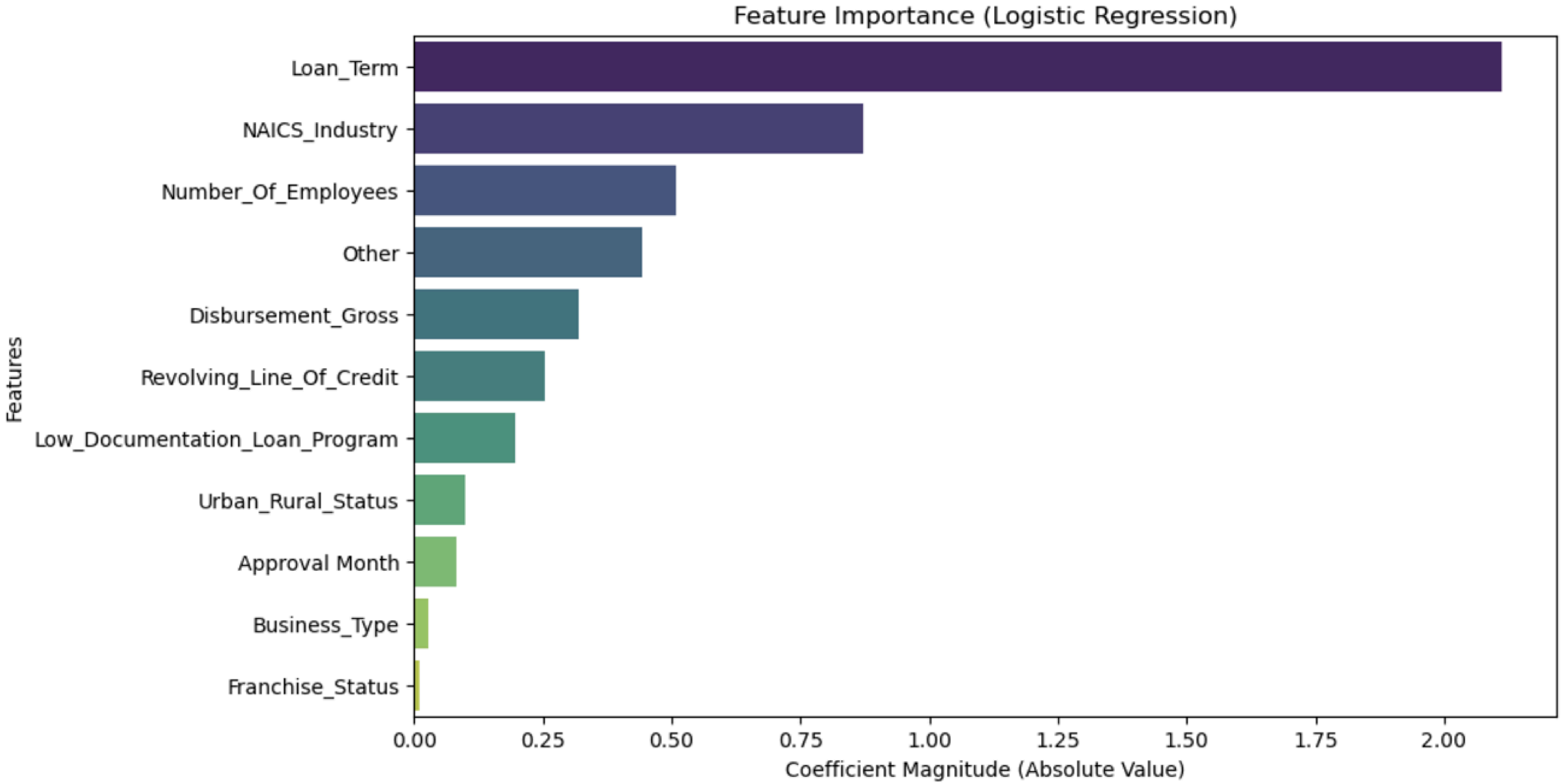


Takeaway

Positive coefficients **boost likelihood** of positive class

Negative coefficient **reduce likelihood** of positive class

Feature Importance



Logistic Regression Model

03

Key Features

1. Loan Term - Months
2. Number of Employees - Number of business employees
3. Revolving Line of Credit - Y/N
4. Low Documentation Loan Program - Y/N
5. Disbursement Gross - Amount disbursed
6. NAICS Codes - Industry Classification Code
7. Franchise Status - Y/N Franchise
8. Month of Approval
9. Business Type - New/Existing
10. Urban Rural Status - Urban/Rural
11. Month of Approval - Month
12. Region

Dummy Variables

```
exclude_columns = [
    'Loan_Term', 'Number_Of_Employees', 'Revolving_Line_Of_Credit', 'Low_Documentation_Loan_Program',
           'Disbursement_Gross', 'Franchise_Status', 'Loan_Status']

#create dummy variables for all cols except those in exclude cols
df_with_dummies = pd.get_dummies(df, columns=[col for col in df.columns if col not in exclude_columns], drop_first=True)
```

```
#convert boolean cols to integer values (0 and 1)
for col in df_with_dummies.select_dtypes(include=['bool']).columns:
    df_with_dummies[col] = df_with_dummies[col].astype(int)
```

1. NAICS Codes
2. Month of Approval
3. Business Type
4. Urban Rural Status
5. Month of Approval
6. Region

Logistic Regression Model

```
y_train_pred = model.predict(X_train)  # Predict on the training set
y_test_pred = model.predict(X_test)  # Predict on the test set

# Apply Z-score normalization (standardization) to the features
scaler = StandardScaler()

# Fit the scaler on the training data and transform both training and test sets
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Adjust decision threshold to be more conservative (> 0.7 = positive instead of .5)
y_train_prob = model.predict_proba(X_train_scaled)[:, 1]  # Probabilities for the positive class
y_test_prob = model.predict_proba(X_test_scaled)[:, 1]
y_train_pred_adjusted = (y_train_prob > 0.7).astype(int)
y_test_pred_adjusted = (y_test_prob > 0.7).astype(int)

# Logistic Regression model
model = LogisticRegression(max_iter=1000, class_weight='balanced', random_state=42)
model.fit(X_train_scaled, y_train)  # Train the model on the scaled training data

# need a low false positive
print(classification_report(y_test, y_pred))
```

Class Weight

Addresses class imbalance, gives more priority to minority class during training, prevents bias

Random State

Ensures reproducibility by controlling randomness in data splitting

Z-Score Normalization

Scales data so all features equally contribute to the model, prevents features with high ranges dominating the learning process

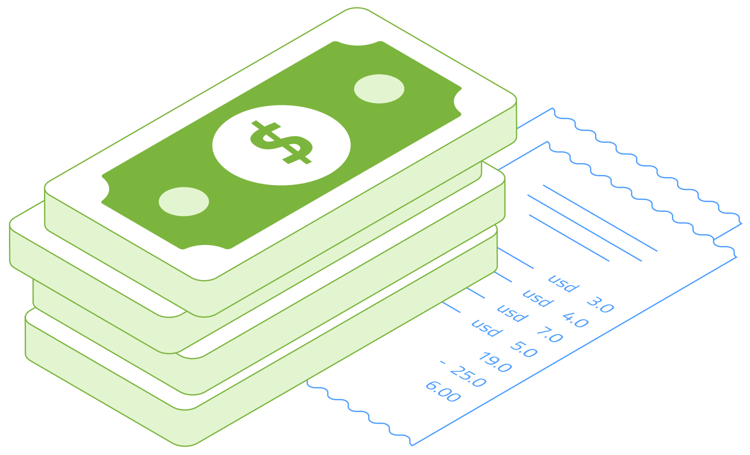
Threshold Adjustment

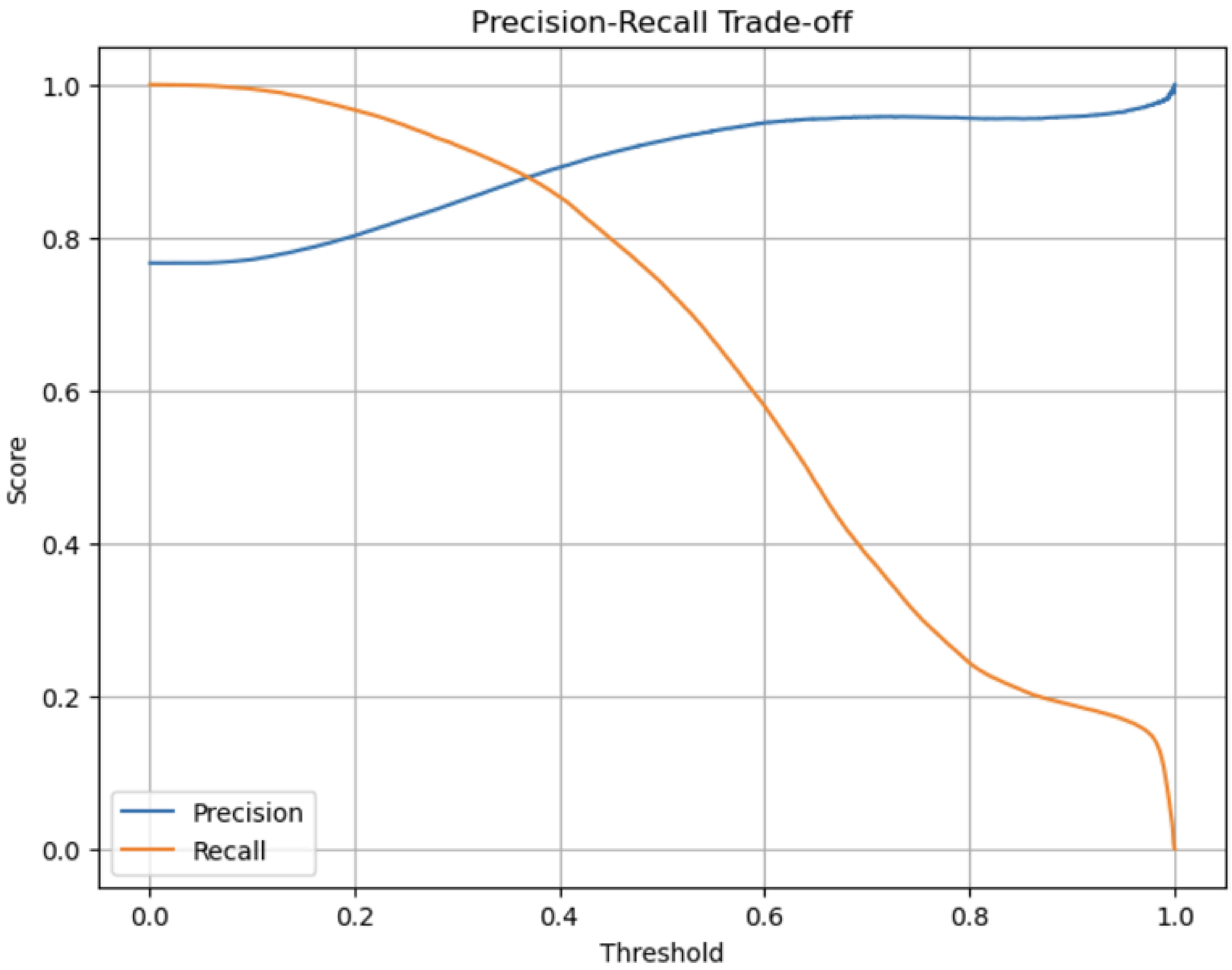
Adjusted from 0.5 to 0.7 to reduce false positives

Evaluation Metrics for LRM

04

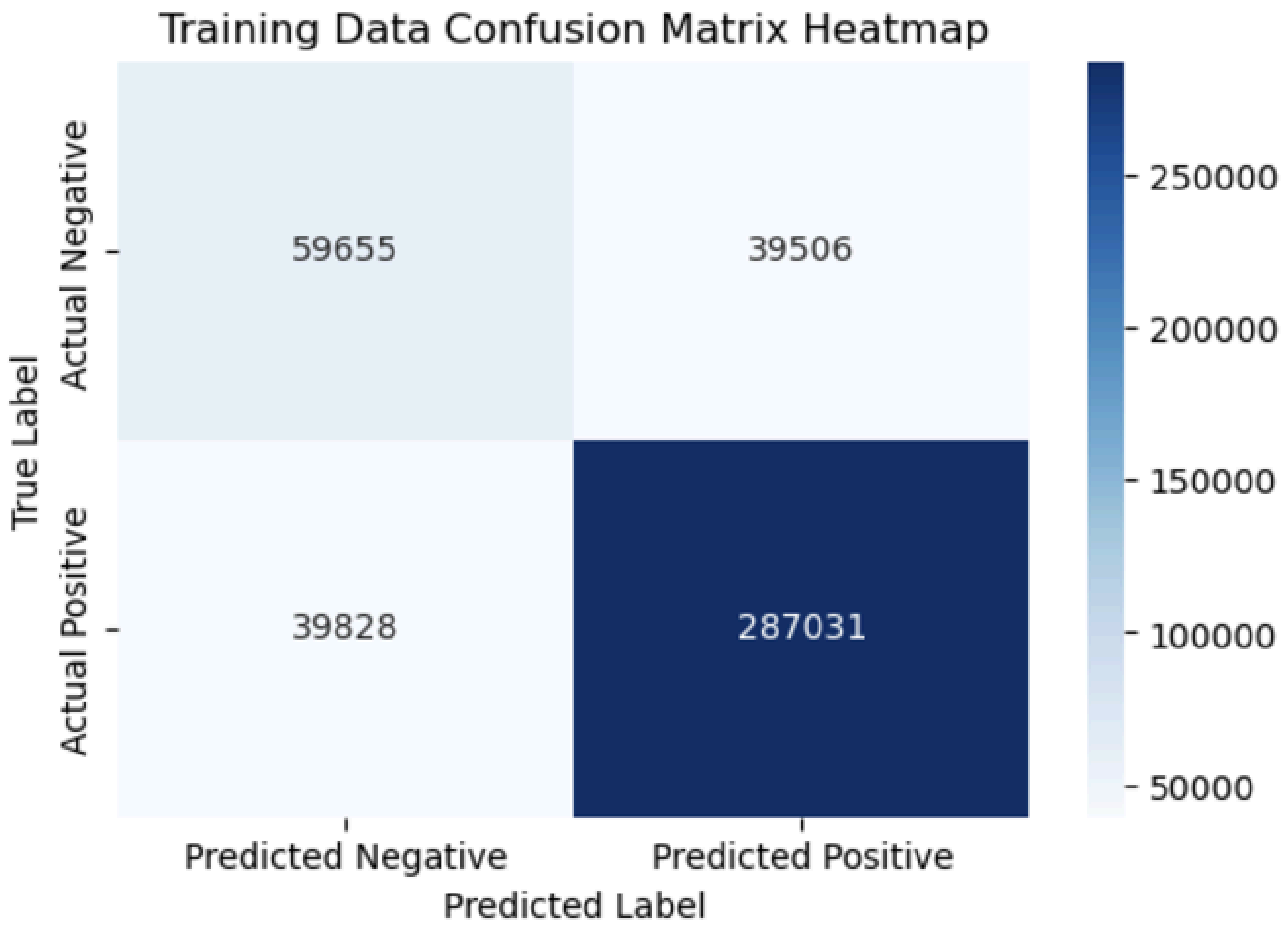
Precision-Recall Curve

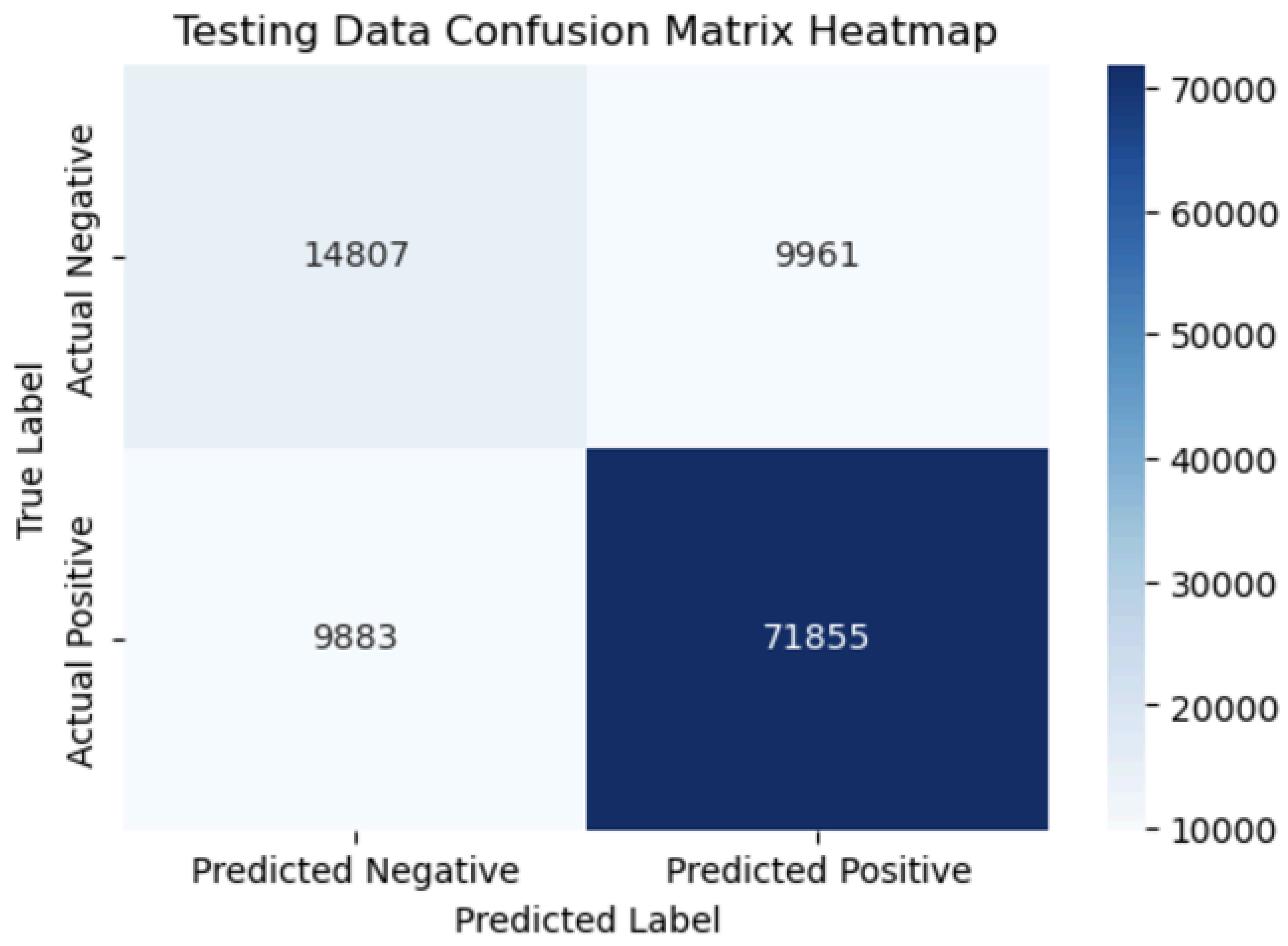




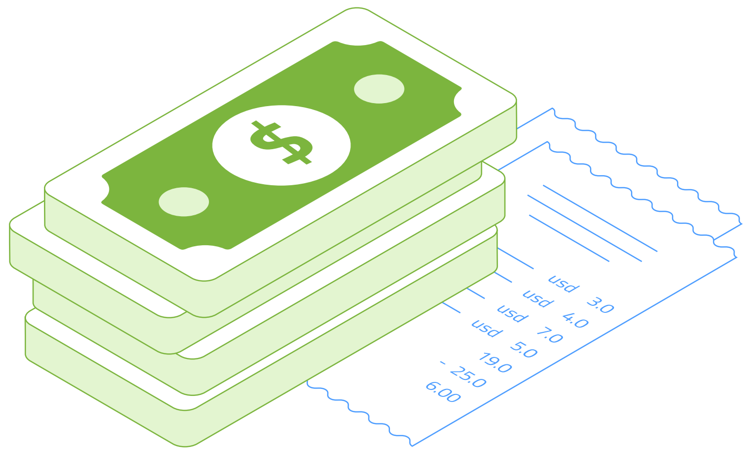
Confusion Matrices

with Precision Recall Optimization





Classification Report



Classification Report for Training Data:

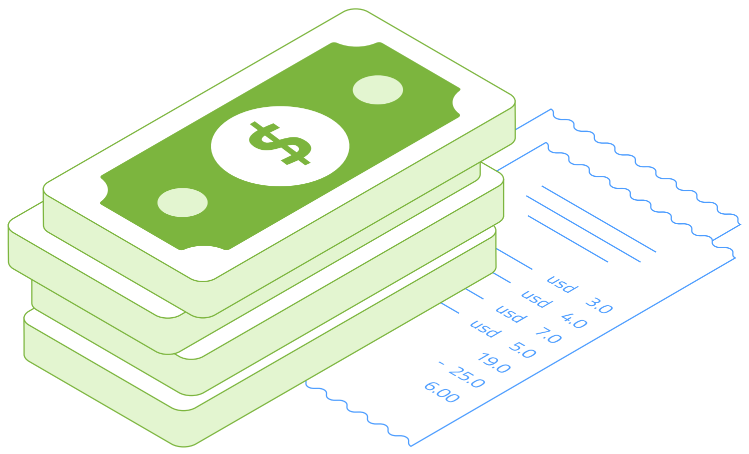
	precision	recall	f1-score	support
0	0.60	0.60	0.60	99161
1	0.88	0.88	0.88	326859
accuracy			0.81	426020
macro avg	0.74	0.74	0.74	426020
weighted avg	0.81	0.81	0.81	426020

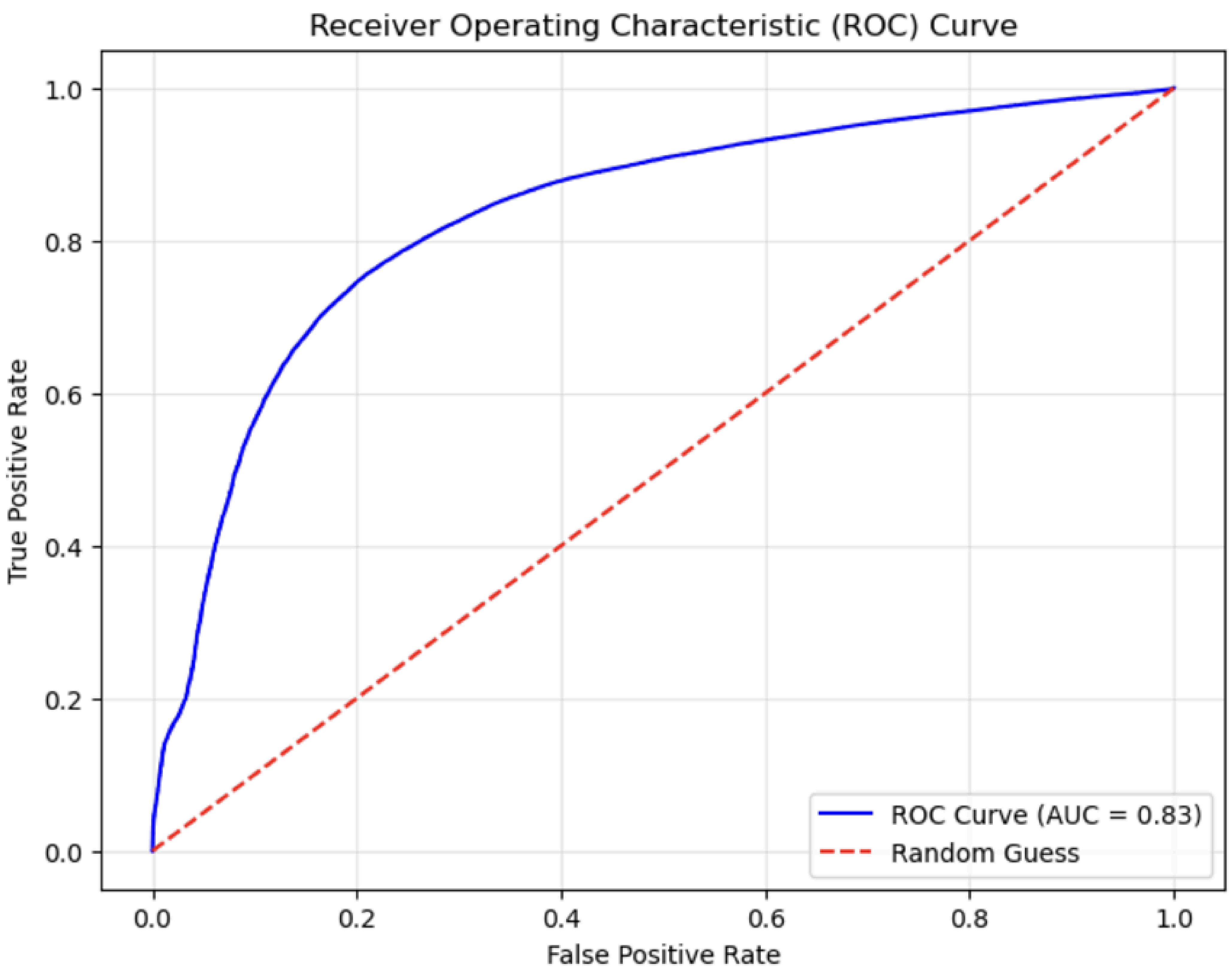
Classification Report for Testing Data:

	precision	recall	f1-score	support
0	0.60	0.60	0.60	24768
1	0.88	0.88	0.88	81738
accuracy			0.81	106506
macro avg	0.74	0.74	0.74	106506
weighted avg	0.81	0.81	0.81	106506

Average Profit for Training Set: 3694.08
Average Profit for Test Set: 3660.77

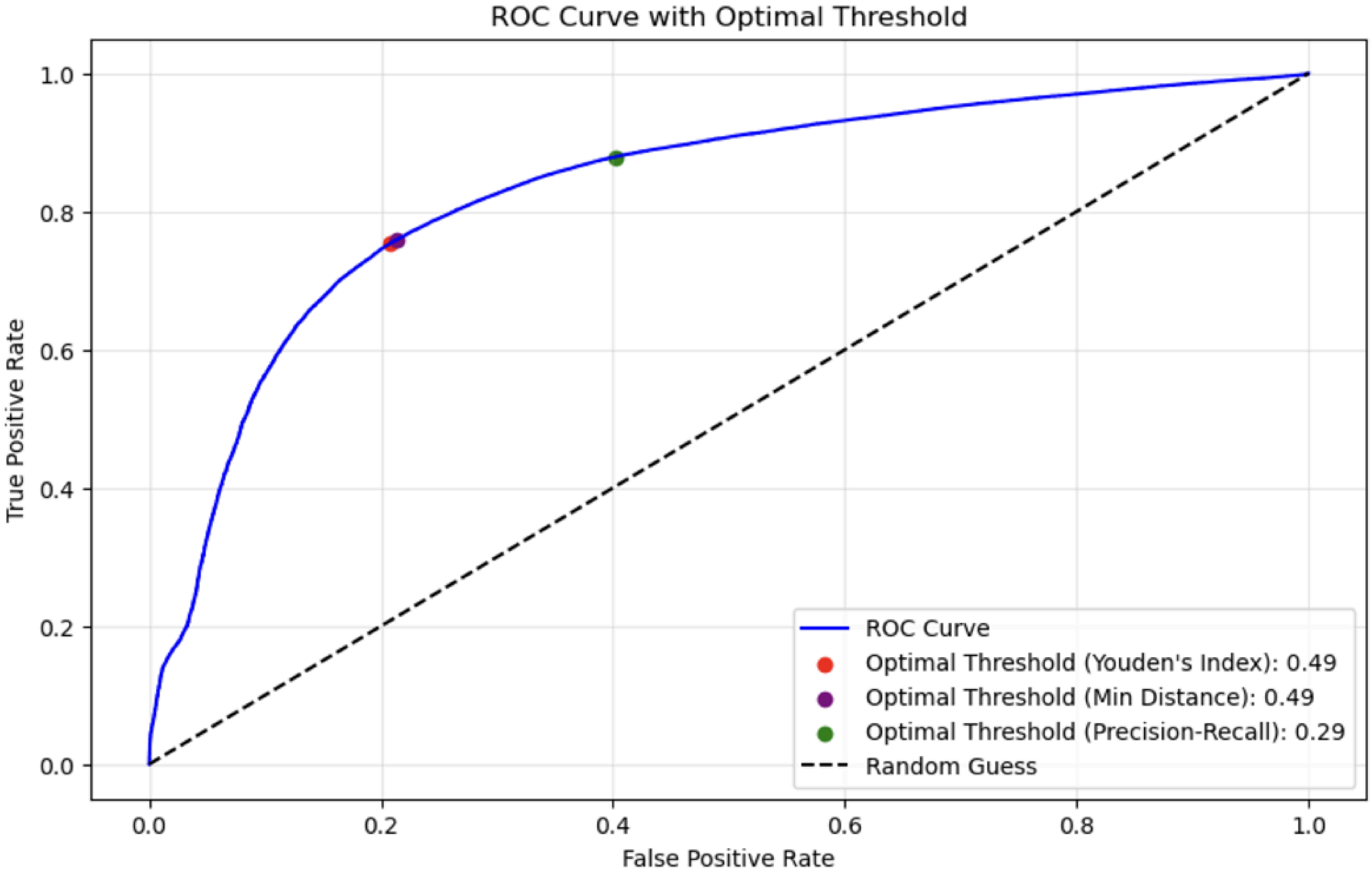
ROC Curve





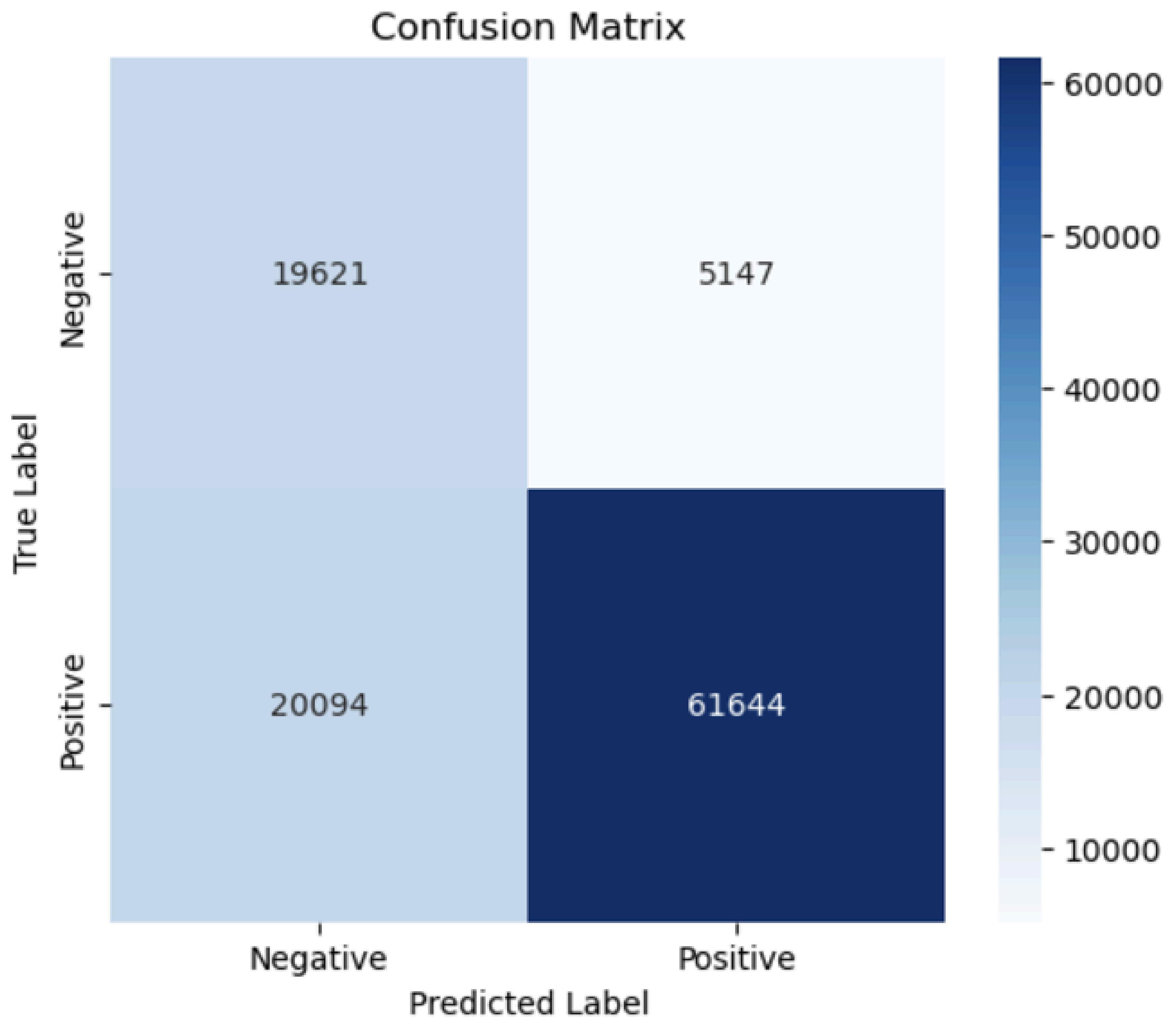
Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (AUC = 0.83)
Random Guess

ROC Curve w/ Optimal Thresholds



Optimal Thresholds:

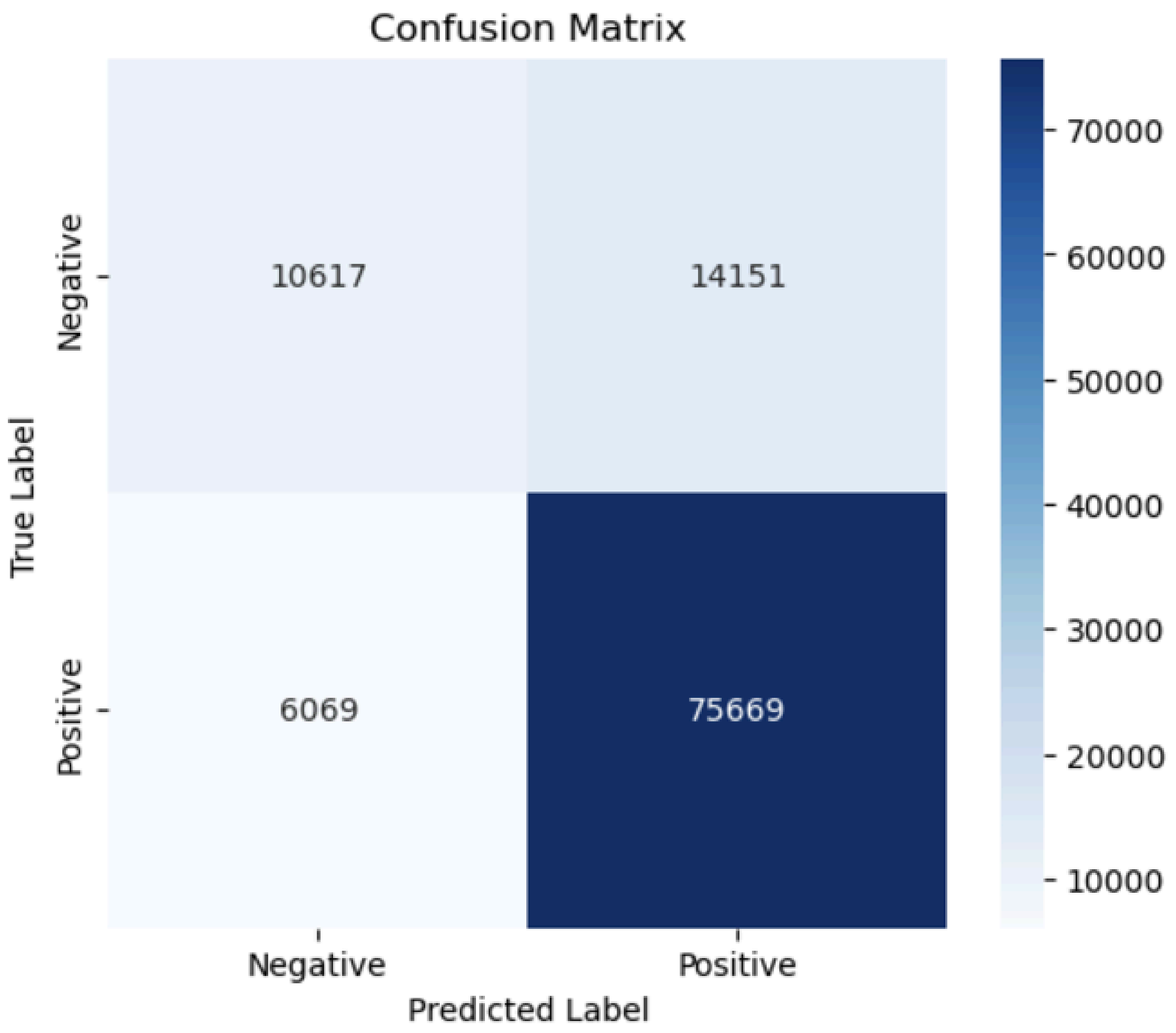
Youden's Index & Min Distance




```
Average Profit for Training Set: 4266.25
Average Profit for Test Set: 4200.46
```

Optimal Thresholds:

ROC Precision Recall




```
Average Profit for Training Set: 4266.25
Average Profit for Test Set: 2981.61
```

Cost-Benefit Analysis

05

Correct Classification Benefit

Denying High-Risk Loans

Avoids financial losses by not approving loans that are likely to default

Approving High-Risk Loans

Ensure the bank earns steady payments and interest income from loans that are likely to be repaid

AVG PROFIT

$4,257.20
Training Set

$4,293.94
Test Set

Represents overall profit/loss and combine correct classification and misclassification between low and high risk loans.

Potential Risks

Financial Loss

Result from approving high-risk loans that default, leading to significant disbursement amounts being lost

Loss of Profit

Occurs when low-risk loans are incorrectly denied, resulting in missed income from loans that would have been repaid

Recommendations

06

Optimal Threshold

To maximize financial gain and minimize risk, we can implement a probability cut-off. This cut off can lead to two possible choices:

Above 0.5 Threshold

- Leads to a higher cutoff
- Results in a safer approach by approving fewer loans
- Reduces the risk of defaults

Below 0.5 Threshold

- Leads to a lower cutoff
- Benefits more small businesses by approving more loans
- Increases the risk of defaults

Integration Plan for Loan Approval Process

Implement Model:

- Include a logistic regression model and apply the chosen probability cut-off.

Provide Training:

- Inform employees about the model and provide training to understand its results.

Record & Monitor Performance:

- Regularly check how well the model is performing.
- Track relevant features that affect the probability of loan approval.

Update Model:

- Collect new data and retrain the model regularly.
- Incorporate suggestions and feedback for continuous improvement.

Taking Action:

- Ensure the model follows all steps and protocols.
- Protect important loan information with robust security measures.

THANK YOU!