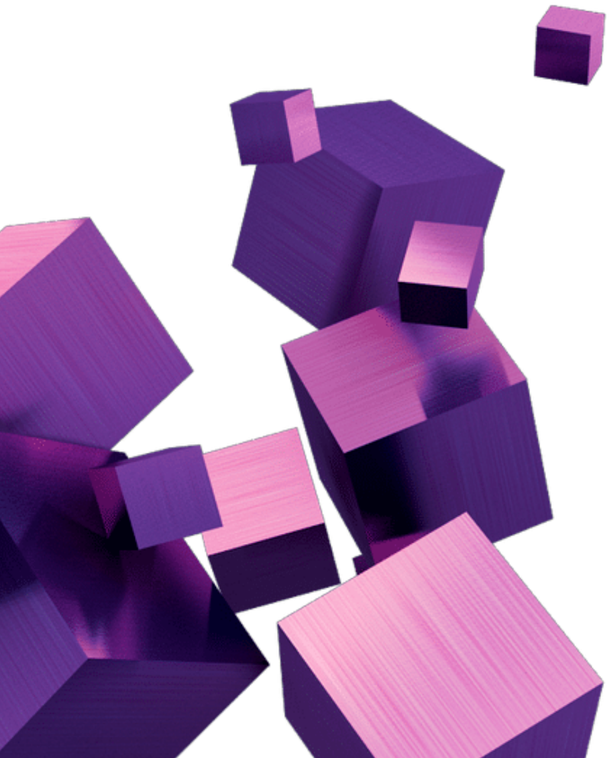
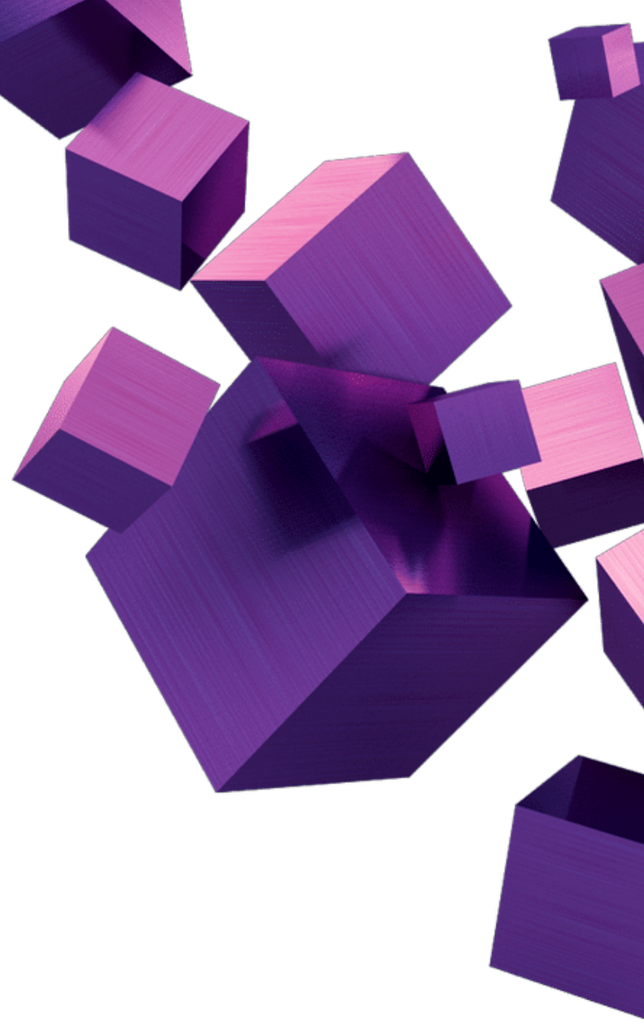
Unsupervised Techniques to Empower Supervised Classification Models

Presentated by Ahmed Taha

Table Of Content

INTRODUCTION





Problem

- The overarching problem faced by the credit card company revolves around balancing **customer base expansion with risk mitigation**.
- The company aims to increase its customer base while ensuring that the risk associated with granting credit to potentially unreliable customers remains manageable.
- However, **the cost of extending credit to a bad customer is significantly higher than the cost of not granting credit to a good customer**, creating a complex financial and operational challenge for the organization.

Financial Challenge

Increased Default Risk

Granting credit to customers with a higher likelihood of defaulting can result in **substantial financial losses** for the company, affecting its **profitability and overall financial health**.

Higher Provisioning Costs

Dealing with bad debts and defaults may necessitate increased provisioning, which can impact the company's reserves and **reduce its ability to invest in growth initiatives or other strategic endeavors**.

Operational Challenge

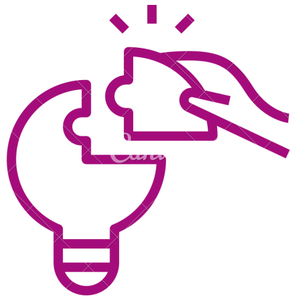
Enhanced Risk Management Complexity

Implementing effective risk management strategies and systems to identify and mitigate potential default risks among the expanding customer base can be operationally challenging, requiring significant resources and expertise.

Efficient Decision-Making Processes

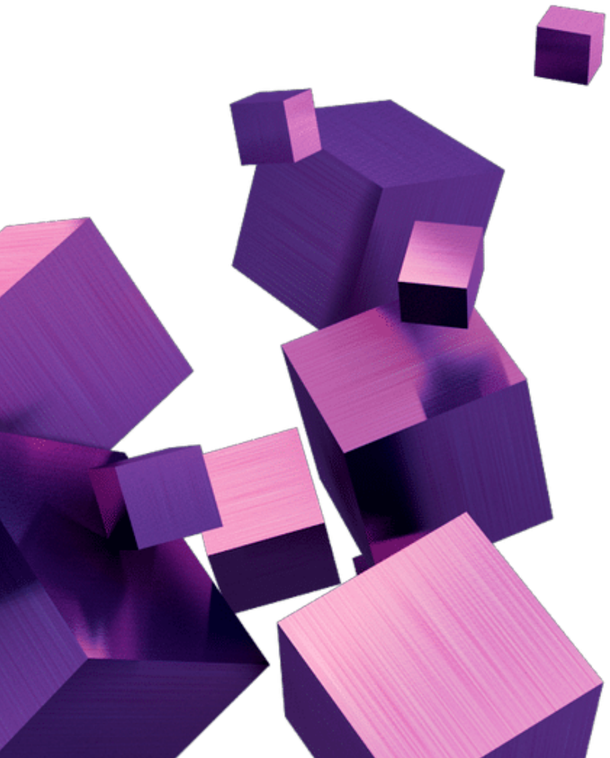
The company needs to make good and fast decisions about who to give credit to, while also thinking about how these decisions will affect their money.

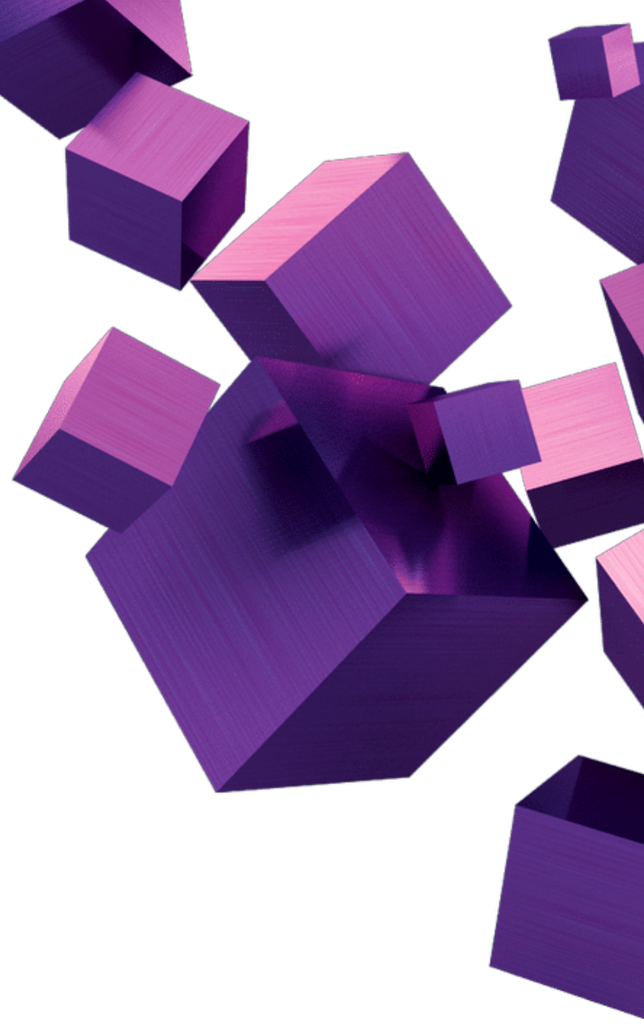
Solution



To address these challenges, the credit card company needs to implement **robust risk assessment methodologies**, including **advanced data analytics** and **machine learning models**, to accurately identify potential default risks.

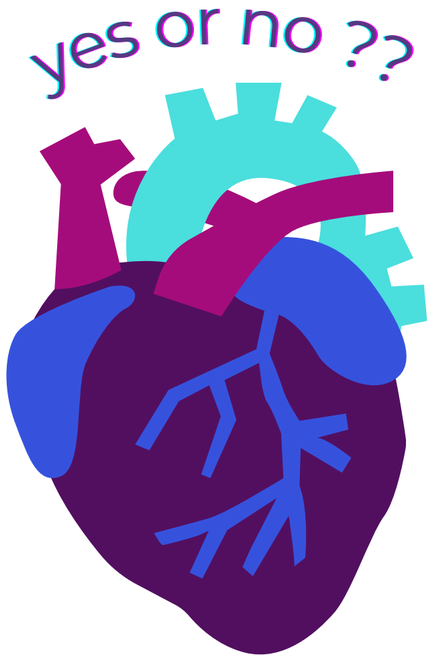
APPROACH





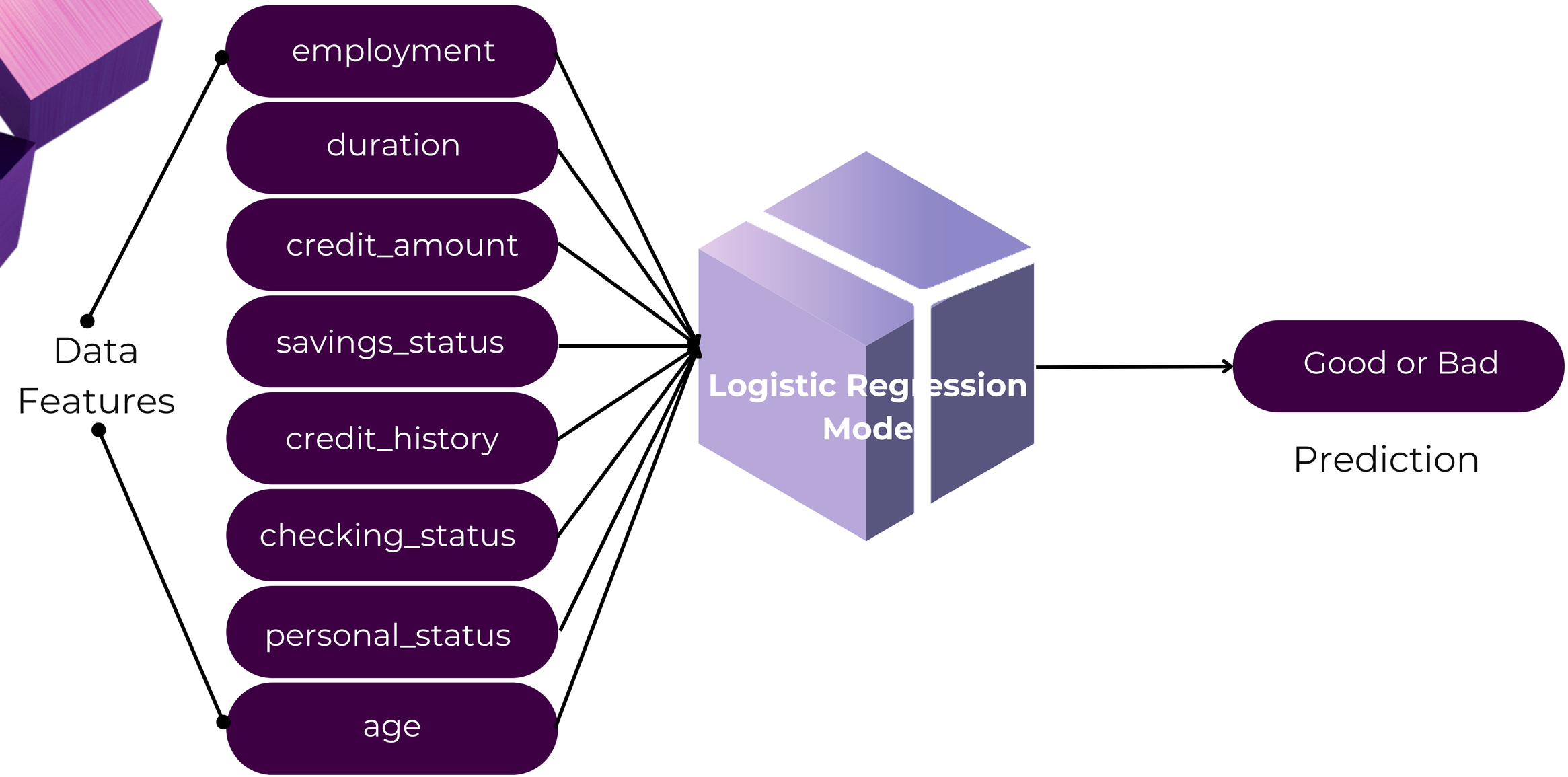
Binary Classification Problems

Whether Person buy or not buy particular product



Whether a disease present or not

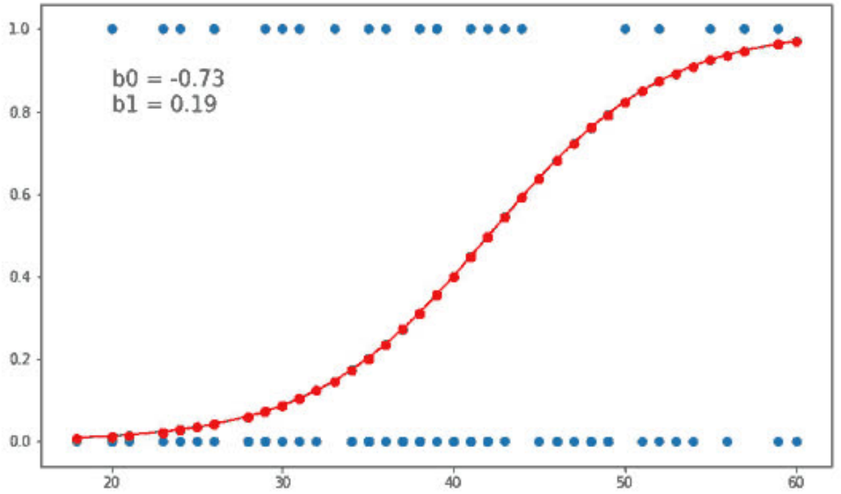
logistic Regression (Base Model)



logistic Regression (Base Model)

How logistic regression work??

In logistic regression, we use a special curve called the **sigmoid curve**, which looks like an **S-shape**. This curve helps us convert the data we have about the customer **into a probability score**, which tells us the **likelihood of them doing the thing we're interested in**. This probability **score is always between 0 and 1**, where 0 means they definitely won't do it, and 1 means they definitely will.



logistic Regression (Base Model)

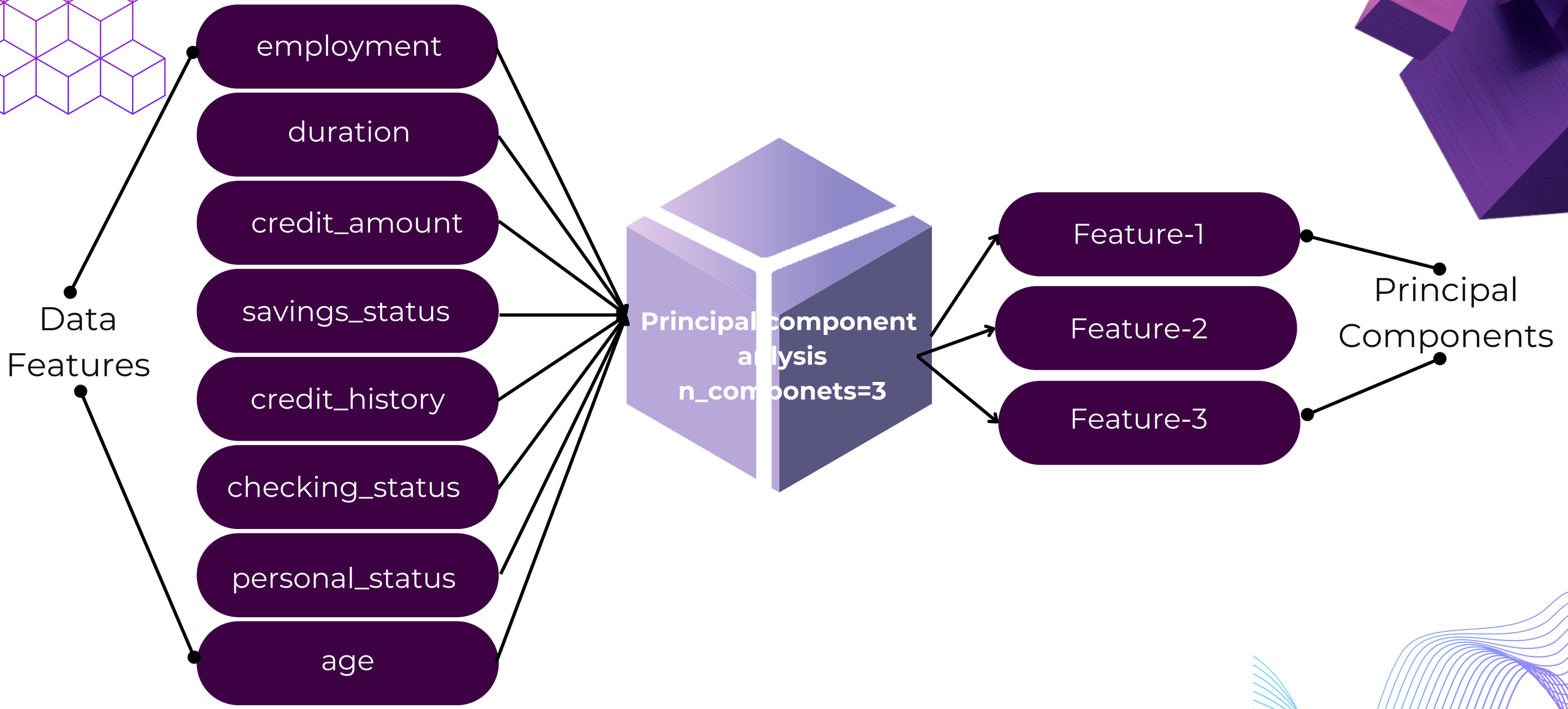
Problems of logistic regression??

- When the relationship between the independent variables and the outcome is highly non-linear and cannot be adequately captured by a linear model.
- When the dataset includes non-independent observations or significant outliers that may impact the model's performance.
- When the problem involves multiple classes and not just binary classification.
- When there is a need for high predictive accuracy in complex datasets, as logistic regression might not perform as well as more sophisticated machine learning algorithms in such cases.

Principal Component analysis

- Principal Component Analysis, or PCA, is a technique that helps us simplify and understand complex data.
- It works like a powerful magnifying glass that allows us to see the most important parts of the data more clearly
- PCA does this by finding new directions in the data that capture the most variation. It then reorganizes the data along these directions, which are called **principal components**
- The great thing about PCA is that it doesn't just throw away information. It keeps the most important parts of the data and discards the less important ones, helping us focus on what really matters

Principal Component analysis



Principal Component analysis

Problems of PCA??

- **Information Loss**: PCA may lead to some loss of information as it compresses the data into fewer dimensions, potentially affecting the interpretability of the results.
- **Non-linear Relationships**: PCA assumes that the relationships in the data are linear, which might not hold true for datasets with complex non-linear relationships.
- **Sensitivity to Outliers**: Outliers can significantly impact the results of PCA, potentially skewing the principal components and leading to misleading interpretations.

AutoEncoder

- An autoencoder is a clever type of artificial neural network that helps us learn **the most important features from complex data**. It works like a smart translator that takes in a lot of information, simplifies it, and then recreates it in a way that's easy to understand.

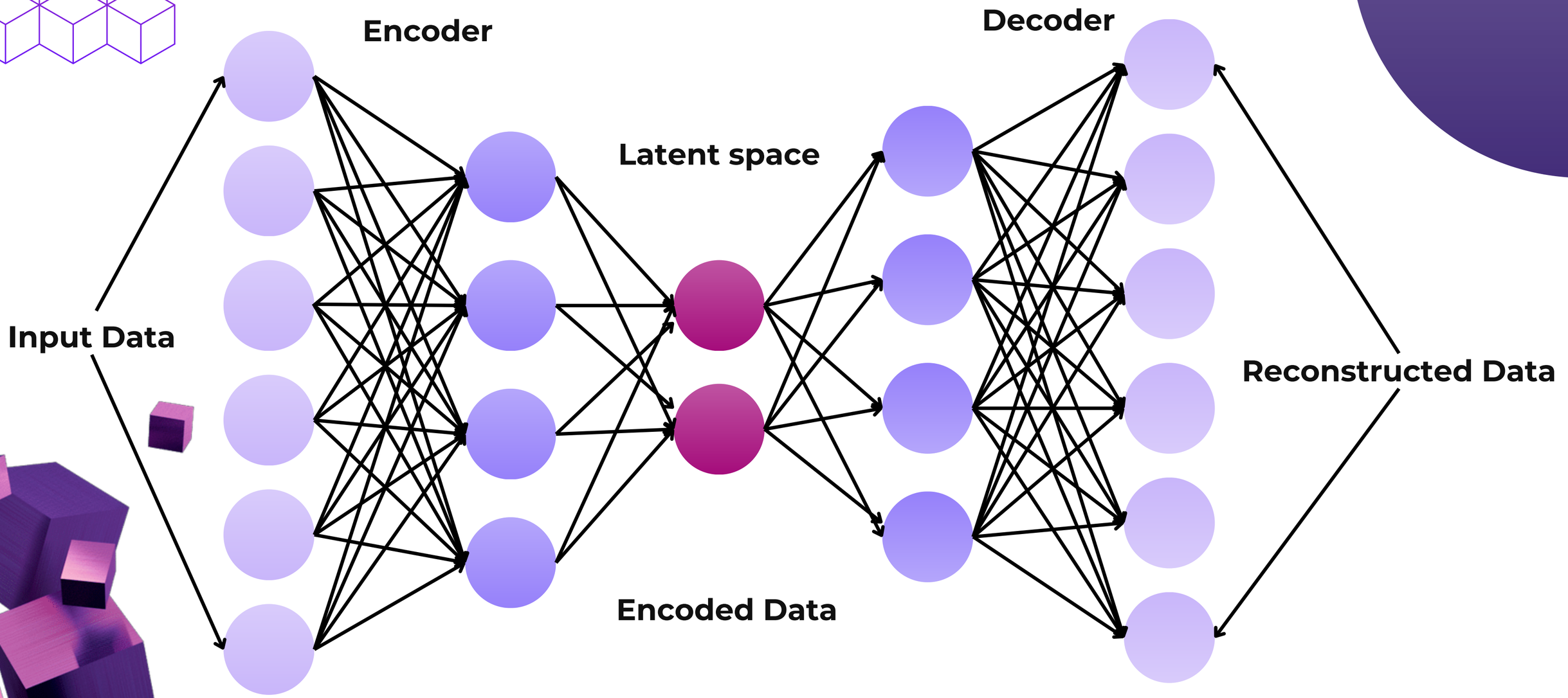
- Autoencoders help **compress features** by learning a condensed representation of the input data through an **encoding process**. This compressed representation, **known as the latent space** or encoding, captures the most essential and salient features of the data.

AutoEncoder

how autoencoders achieve feature compression??

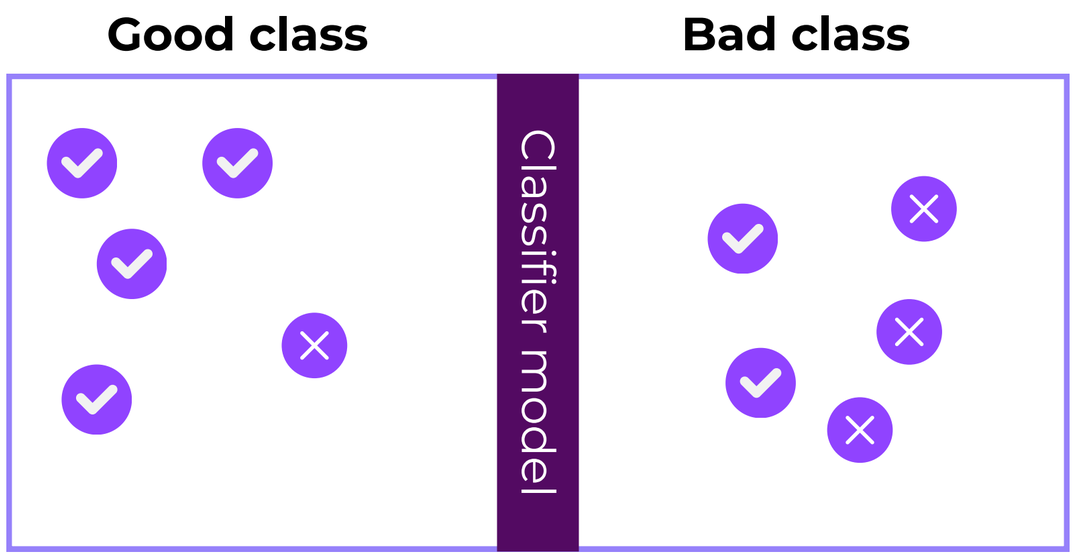
- **Encoding**: The autoencoder's encoder network takes the input data and maps it to a lower-dimensional space, capturing the most important features.
- **Bottleneck Layer**: The bottleneck layer within the autoencoder serves as the bridge between the encoder and decoder. It contains the most compact representation of the input data, where the essential features are preserved while removing redundant or less critical information.
- **Decoding**: The decoder network then reconstructs the compressed representation back into the original data space, aiming to produce an output that closely resembles the initial input.
- **This compressed representation can then be used for various purposes including data visualization, data compression.**

AutoEncoder



Evalution Metrix

- **Precision**: Precision is the measure of how many selected items are relevant. It tells us how often the model is correct when it predicts something.
- **Recall**: Recall is the measure of how many relevant items were selected. It tells us how well the model finds all the relevant cases.
- **F1 Score:** F1 score is a combined metric of precision and recall. It gives us a single number that represents the balance between precision and recall. It's like a way to see how well the model is doing overall, considering both its accuracy and its ability to find relevant cases.



Metrics

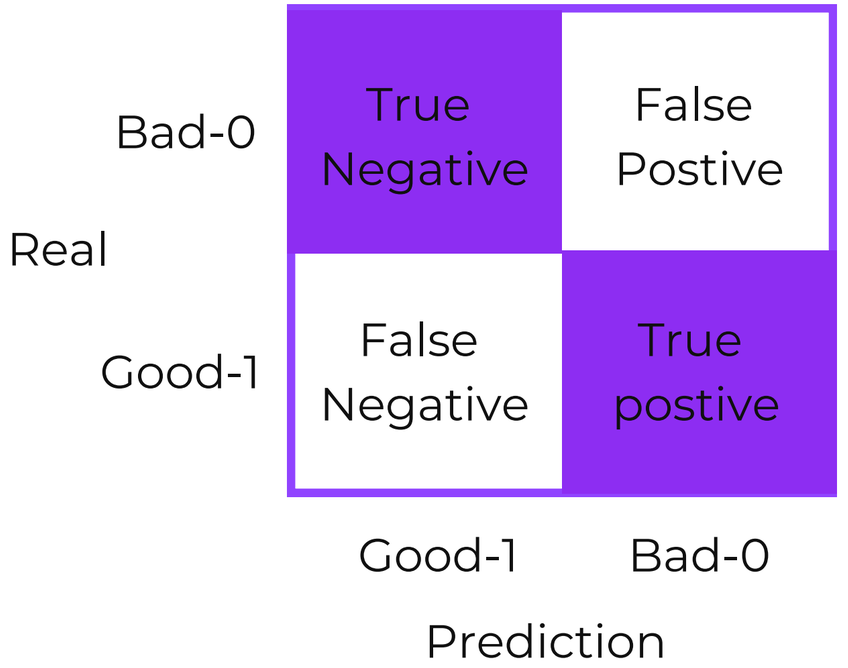
$$\text{Precison(Good Class)} = \frac{\text{✓ ✓ ✓}}{\text{✓ ✓ ✓ ✗}} = \frac{3}{4} = 75\%$$

$$\text{Recal(Good Class)} = \frac{\text{✓ ✓ ✓}}{\text{✓ ✓ ✓ ✓ ✓}} = \frac{3}{5} = 60\%$$

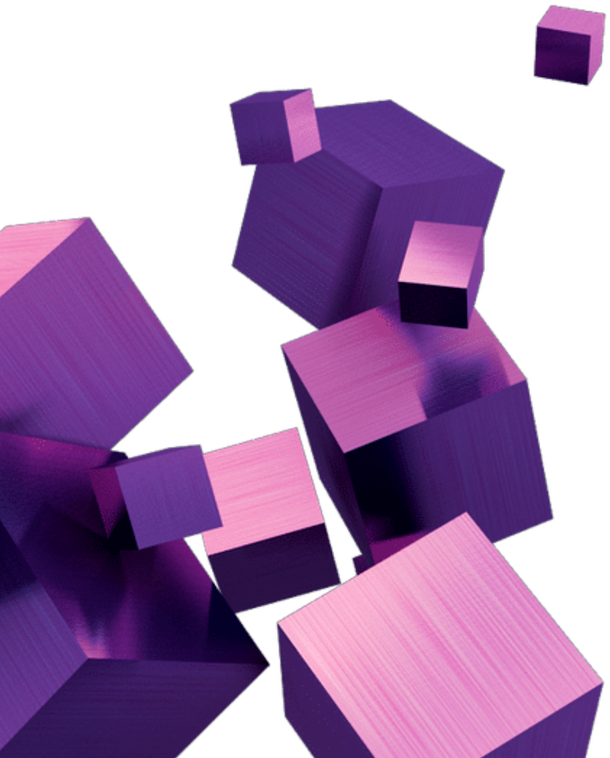
$$\text{f1-score(Good Class)} = \frac{2 \times \frac{3}{4} \times \frac{3}{5}}{\frac{3}{4} + \frac{3}{5}} = 66.7\%$$

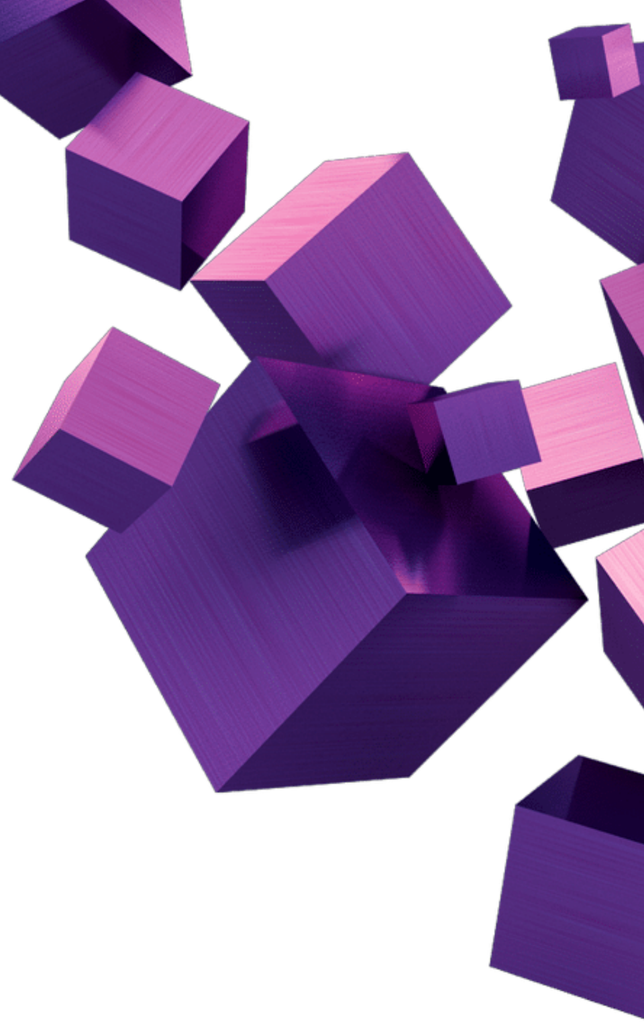
Confusion Metrix

- A confusion matrix is a table that is often used to describe the performance of a classification model.
- It provides a summary of the predictions made by a classification model, comparing them to the actual labels.
- The confusion matrix has four main components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).



Data Analysis

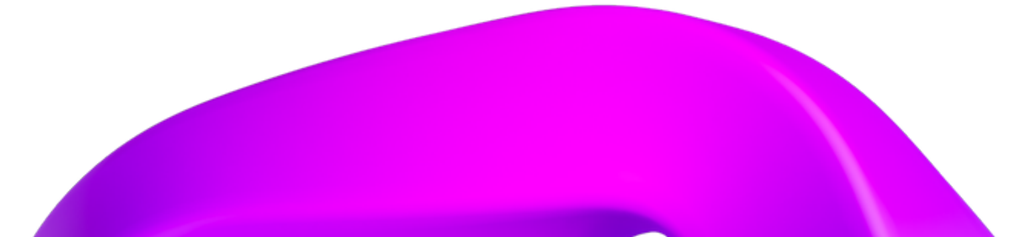




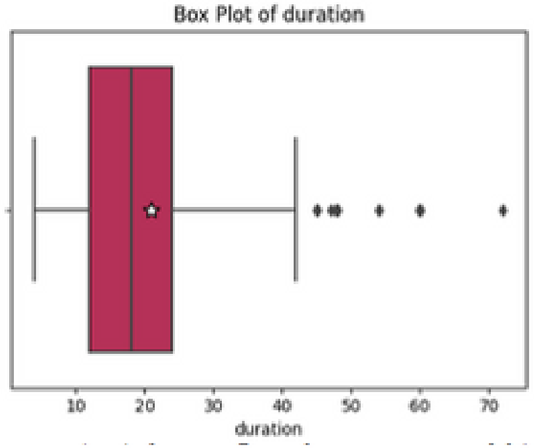
Credit Applicant Data Overview

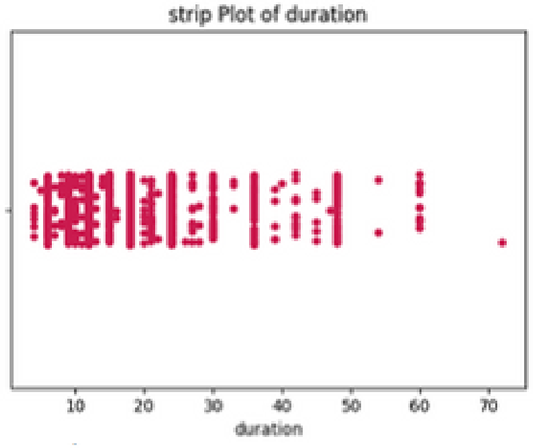
- **Data Size**: 1000 rows, 20 columns
- **Data Attributes**:

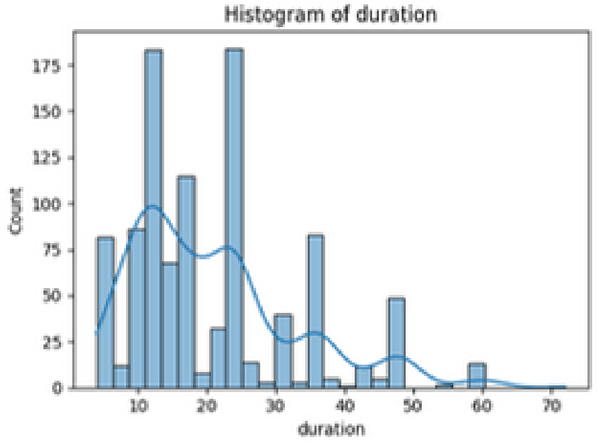
1. **Demographics**: Age, gender, number of dependents, foreign worker status
2. **Financial Information**: Checking account status, credit history, savings status
3. **Credit Details**: Credit duration, amount, purpose, existing credits
4. **Employment**: Employment duration, job type, installment commitment
5. **Residence**: Duration of current residence, housing status, property magnitude
6. **Communication**: Telephone ownership, other payment plans
7. **Classification**: 'Good' or 'Bad' credit class

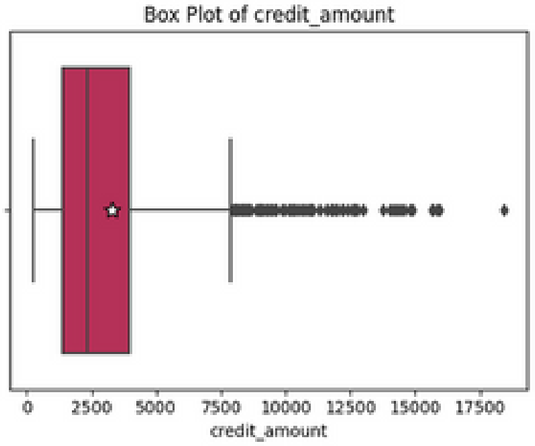


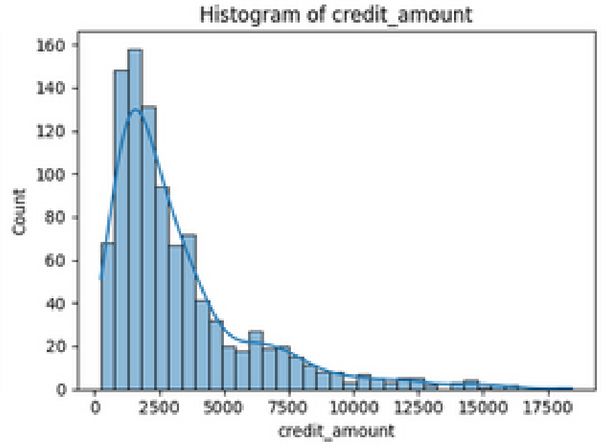
Numerical Features Analysiss

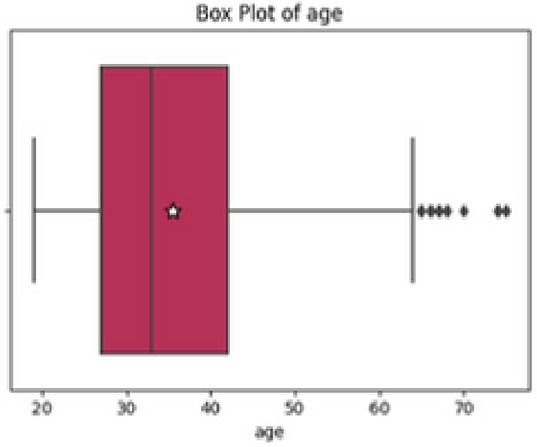


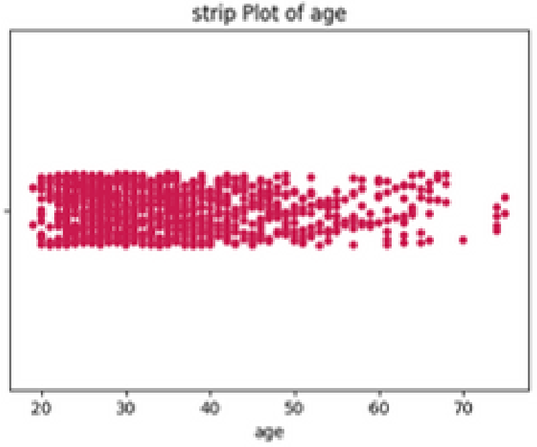


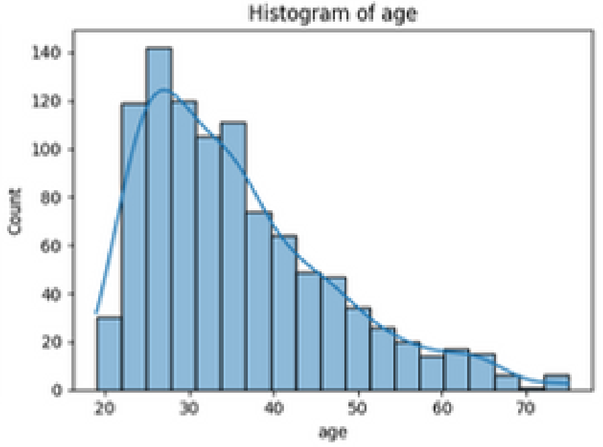




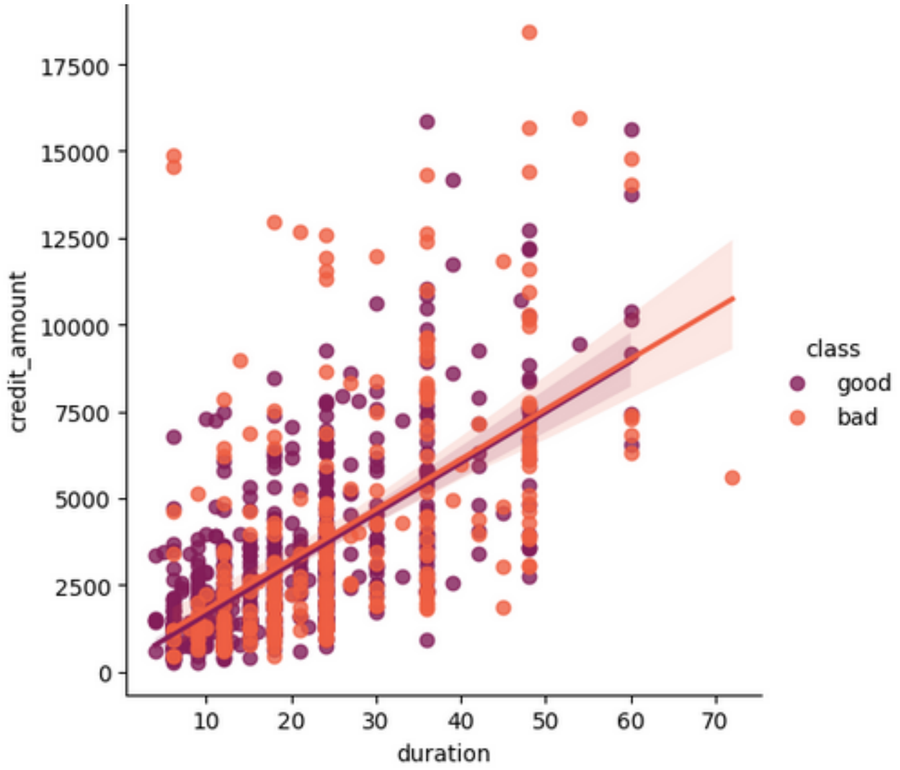


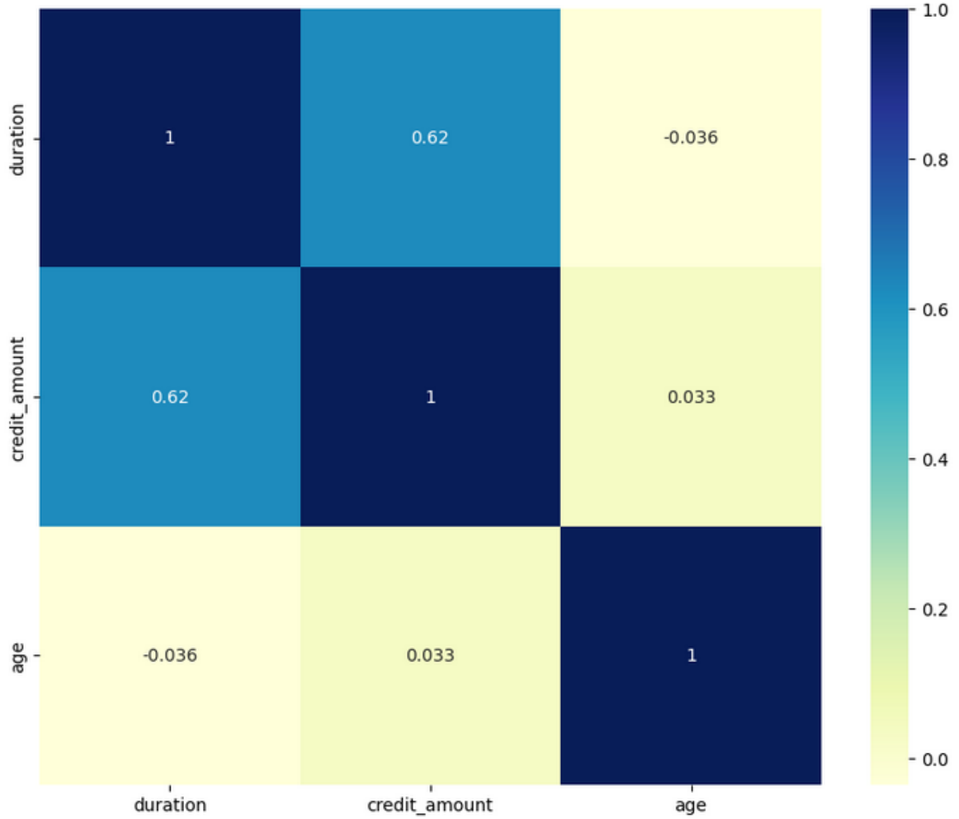




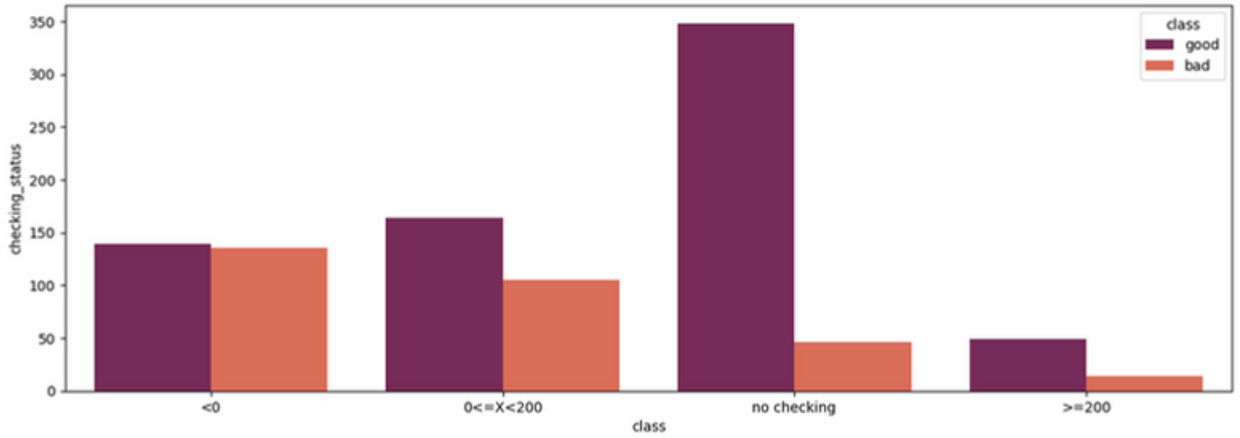


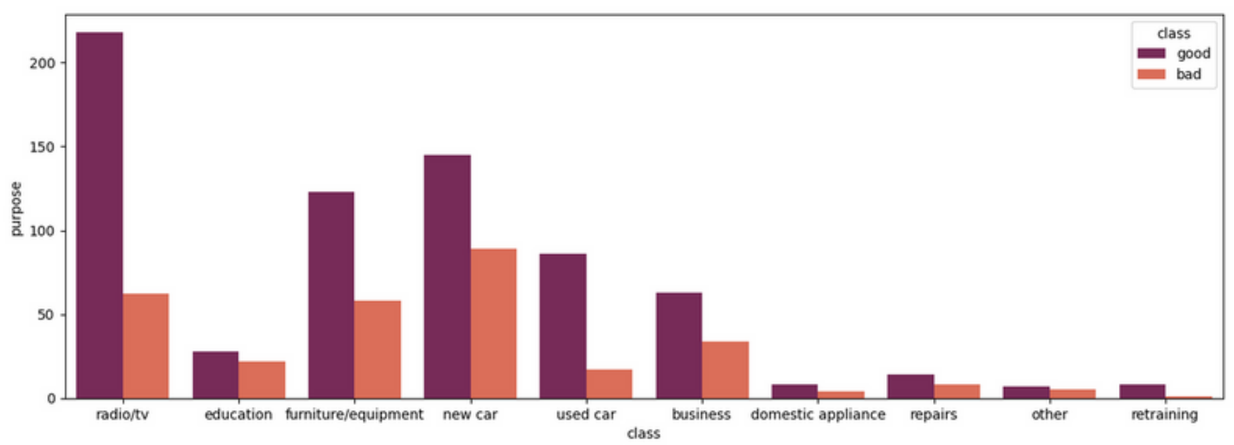
Numerical Features Analysiss Correlation

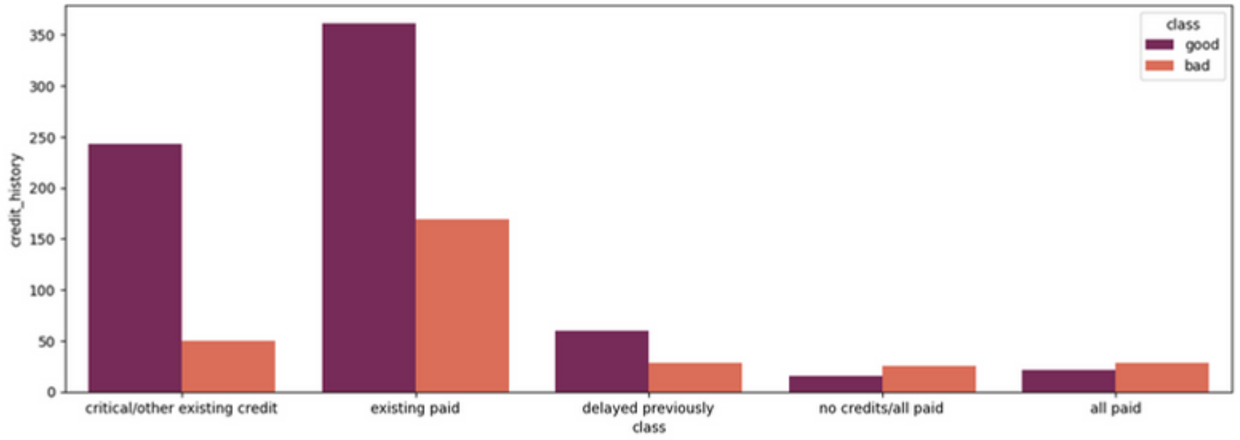


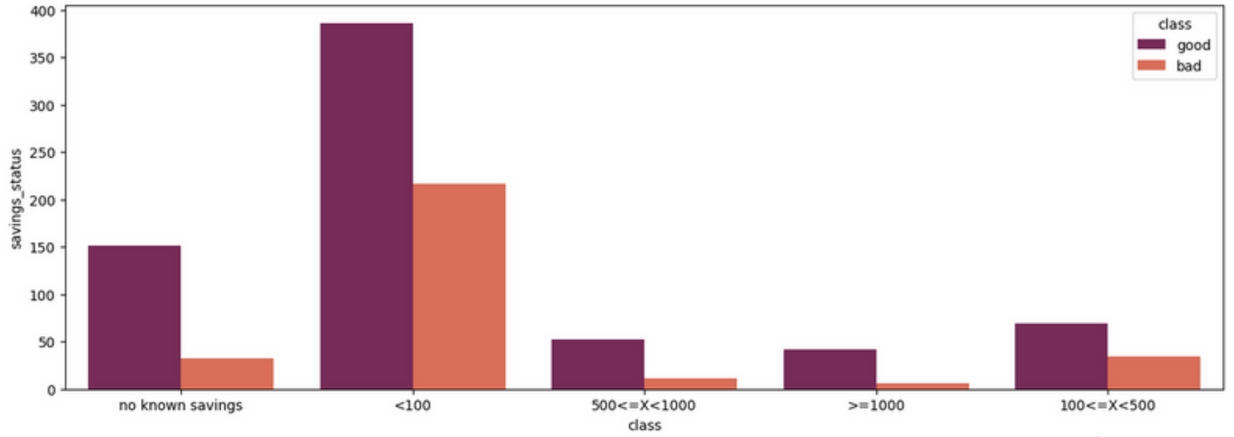


Categorical Features Analysis

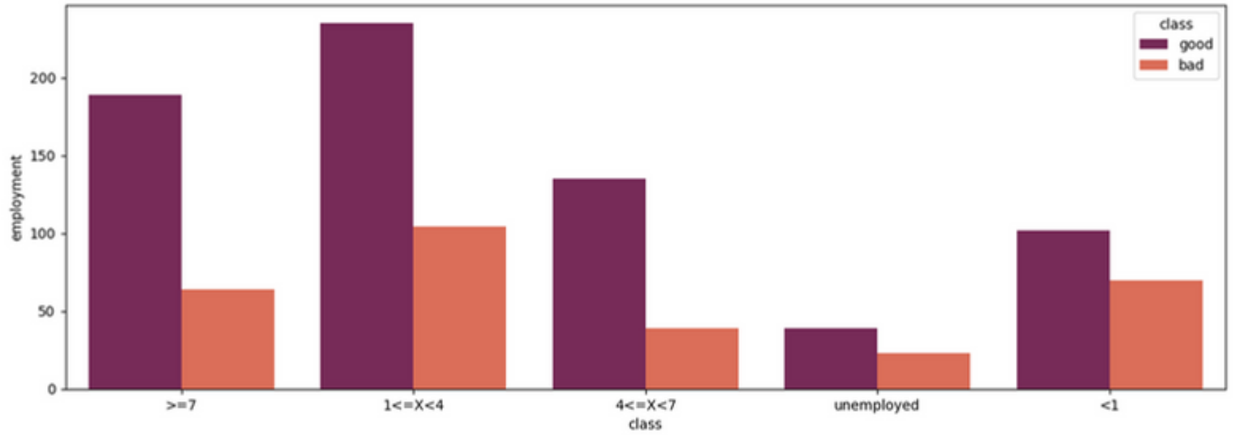


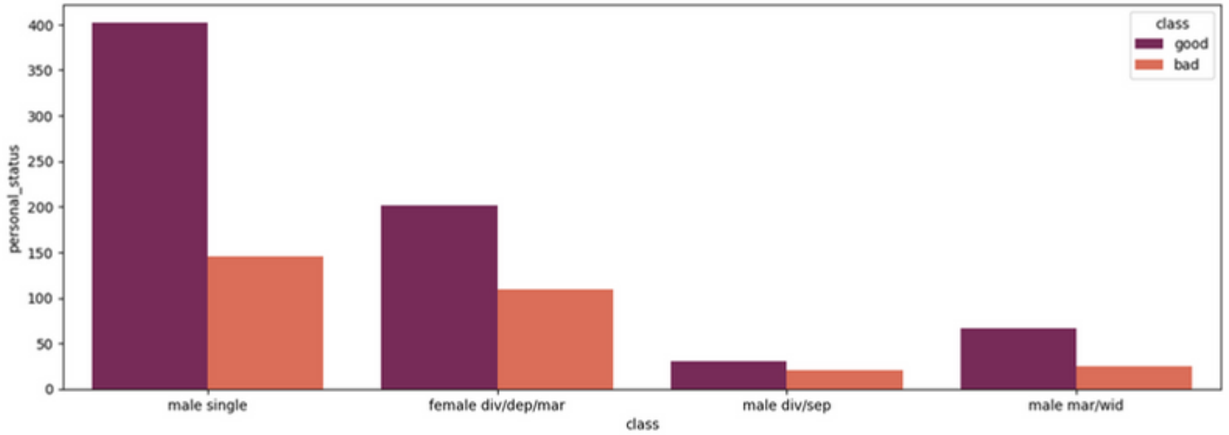


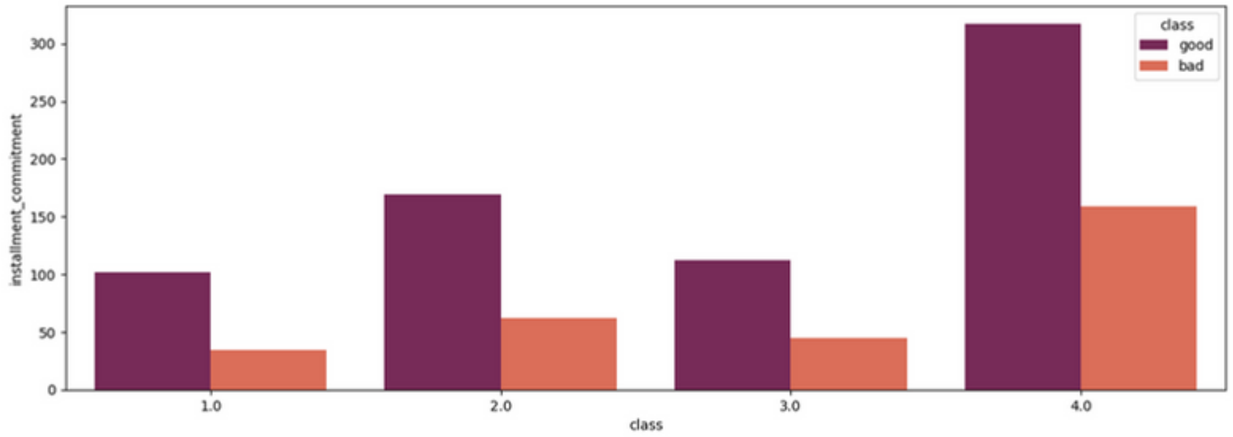


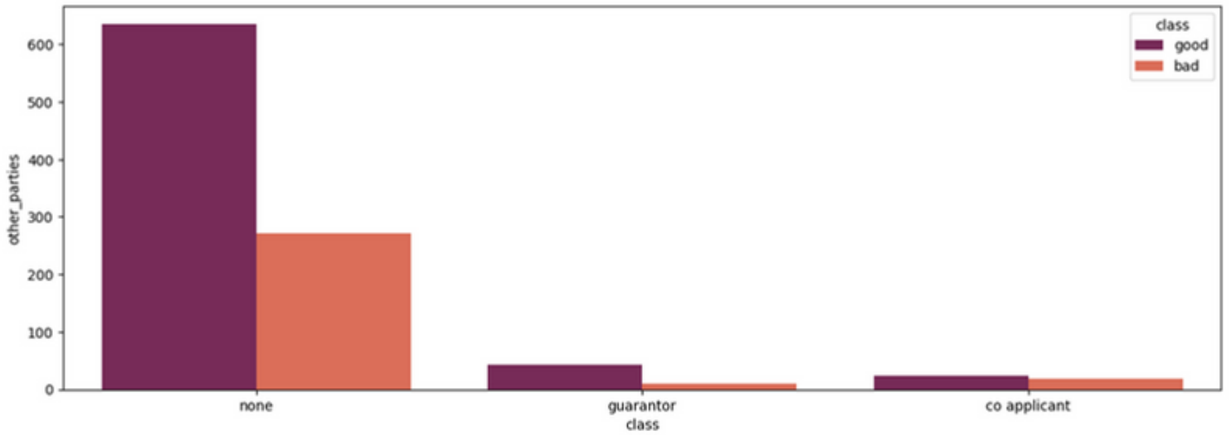


Categorical Features Analysis

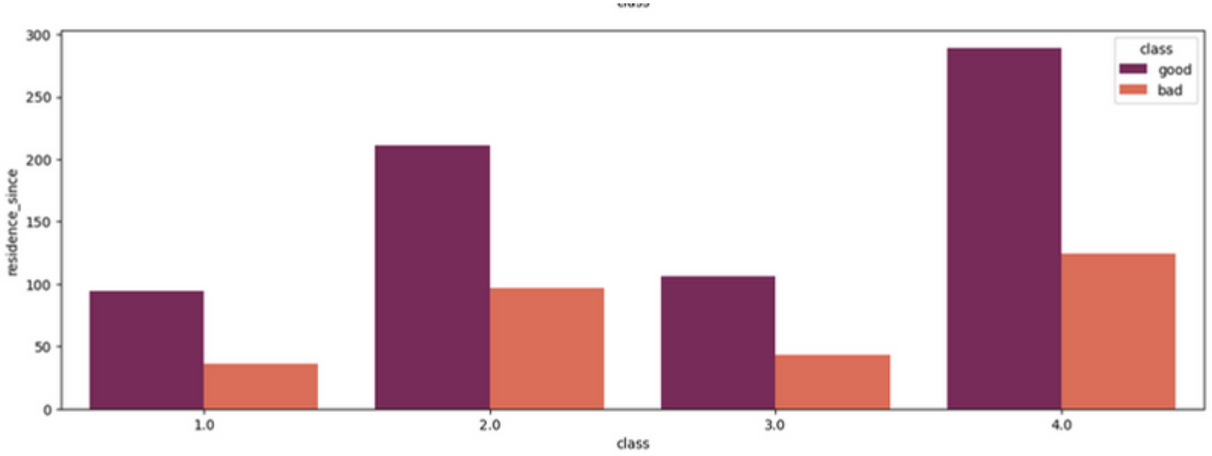


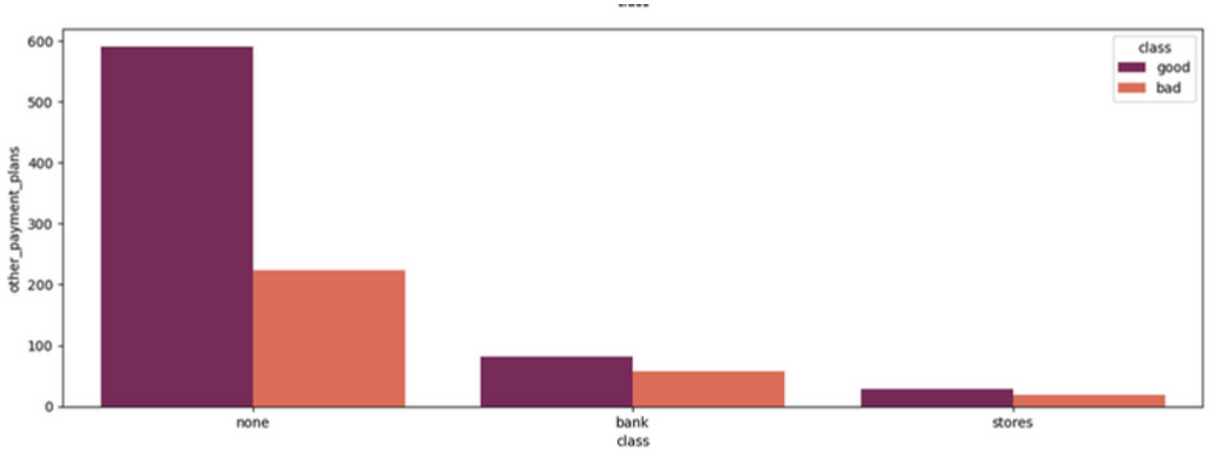


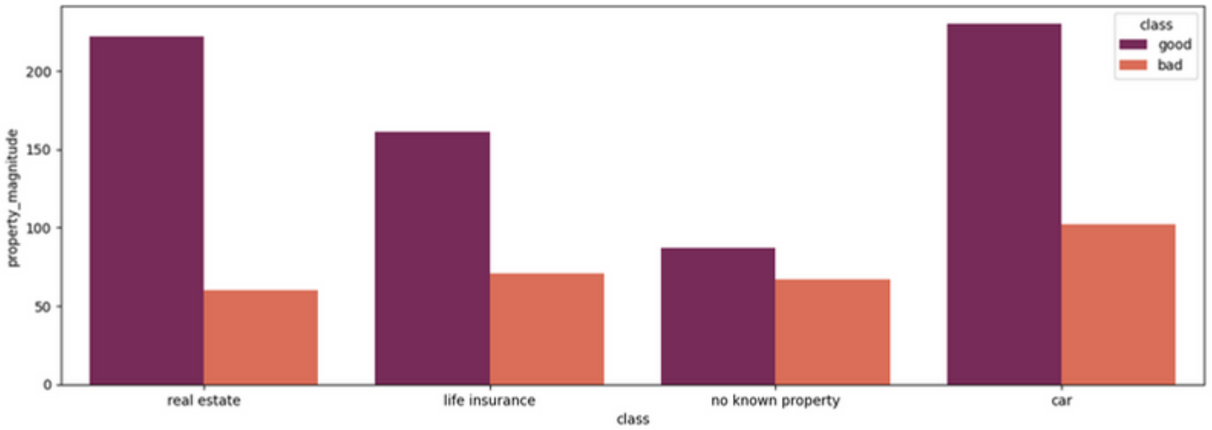


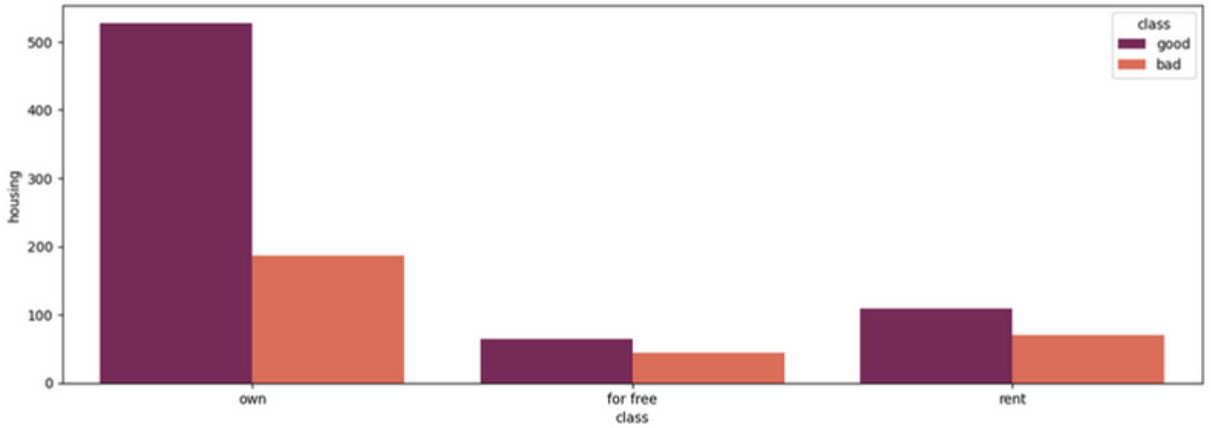


Categorical Features Analysis

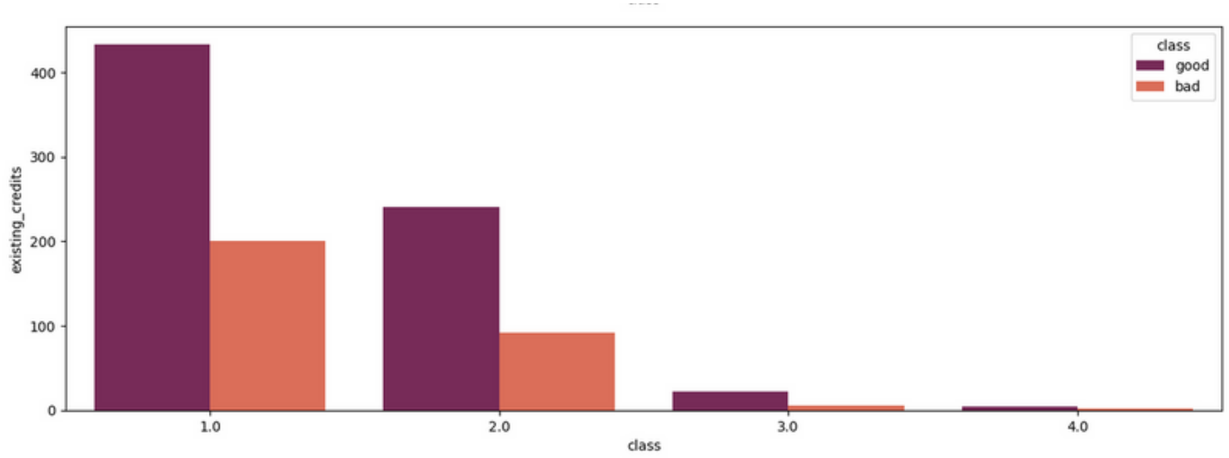


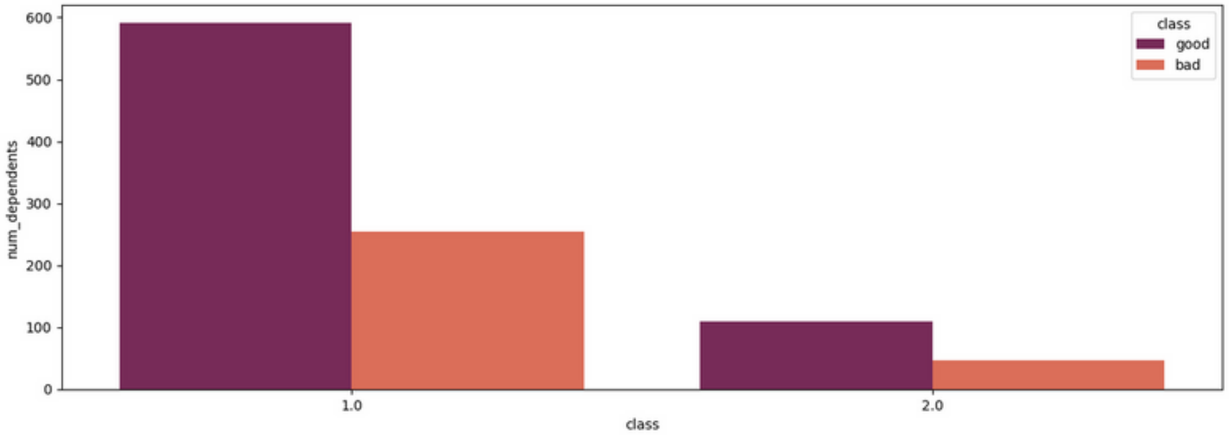


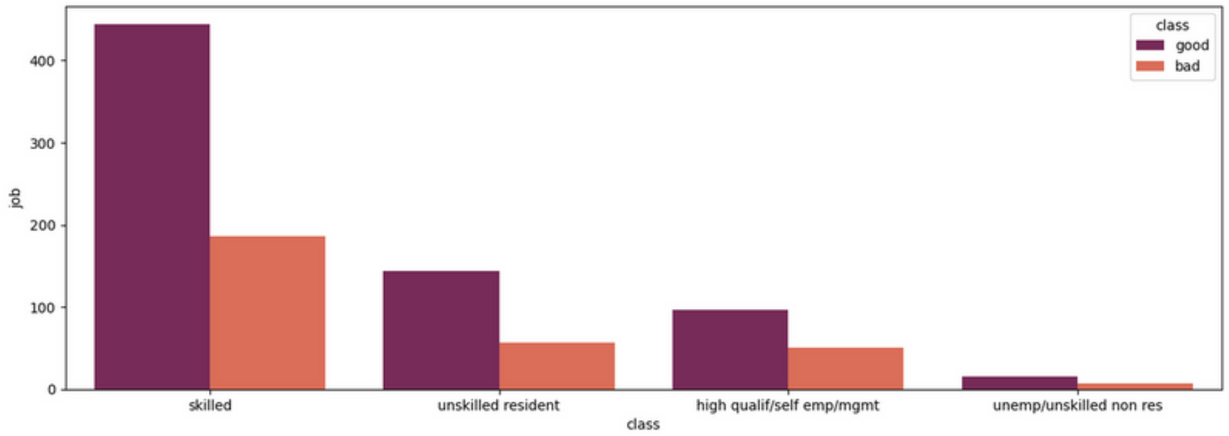


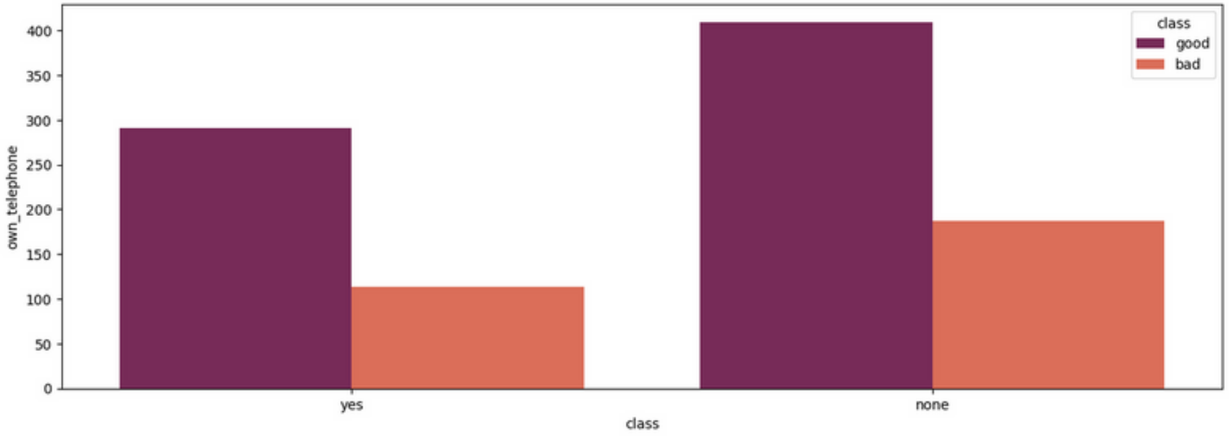


Categorical Features Analysis

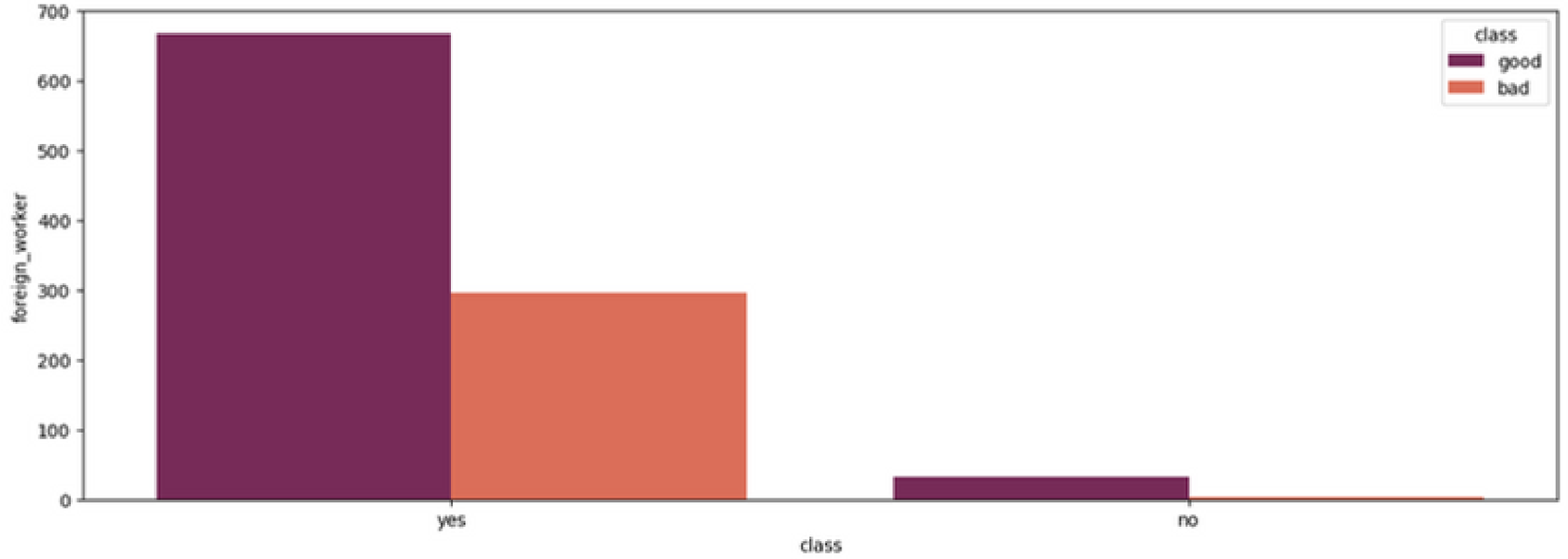


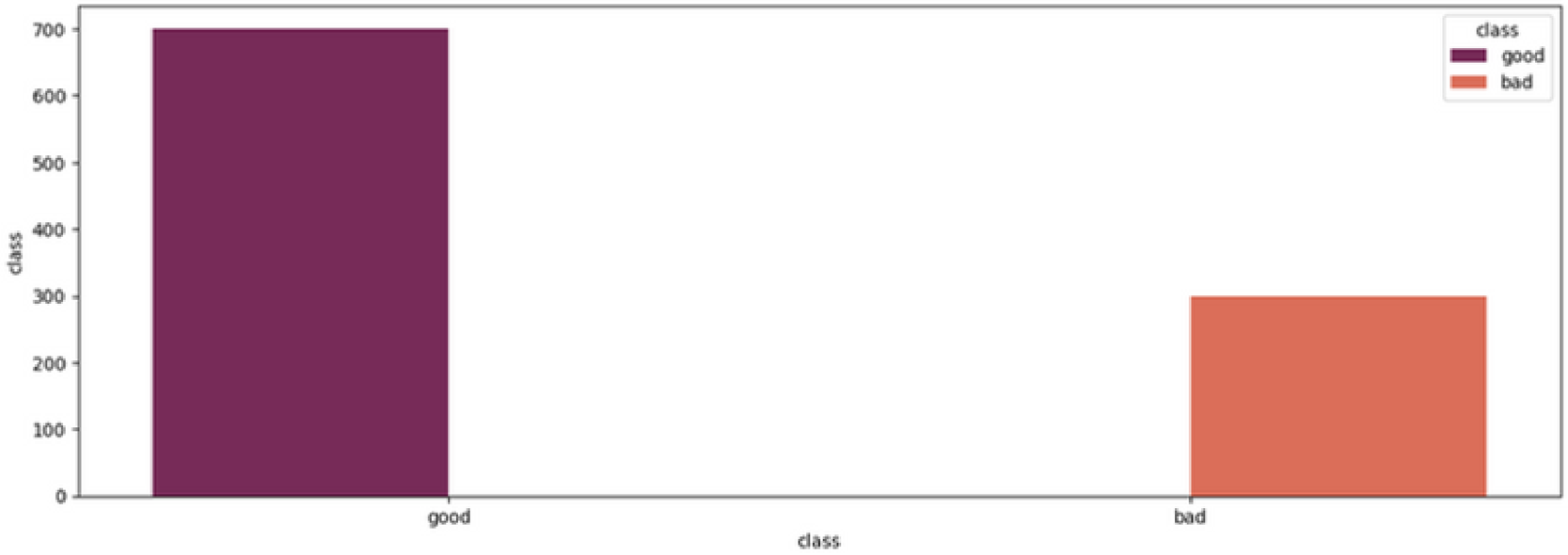






Categorical Features Analysis





Feature selection

Numerical Features

	Feature	Pvalue(Anova)	Pvalue(T-Test)	Result
0	duration	6.488050e-12	6.488050e-12	Accepted
1	credit_amount	8.797572e-07	8.797572e-07	Accepted
2	age	3.925339e-03	3.925339e-03	Accepted

Categorical Features

	Feature	Pvalue(Chi-Squared)	Result
0	checking_status	1.218902e-26	Accepted
1	credit_history	1.279187e-12	Accepted
2	purpose	1.157491e-04	Accepted
3	savings_status	2.761214e-07	Accepted
4	employment	1.045452e-03	Accepted
5	installment_commitment	1.400333e-01	Not-Accepted
6	personal_status	2.223801e-02	Accepted
7	other_parties	3.605595e-02	Accepted
8	residence_since	8.615521e-01	Not-Accepted
9	property_magnitude	2.858442e-05	Accepted
10	other_payment_plans	1.629318e-03	Accepted
11	housing	1.116747e-04	Accepted
12	existing_credits	4.451441e-01	Not-Accepted
13	job	5.965816e-01	Not-Accepted
14	num_dependents	1.000000e+00	Not-Accepted
15	own_telephone	2.788762e-01	Not-Accepted
16	foreign_worker	1.583075e-02	Accepted

Data Cleaning & Preprocssing

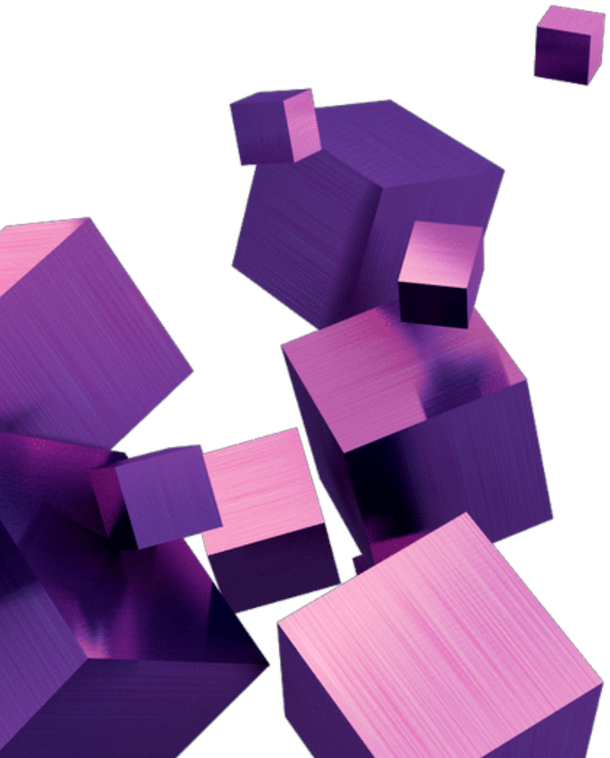
- No missing values
- No duplicates
- Handling the categorical data using **label encoder**

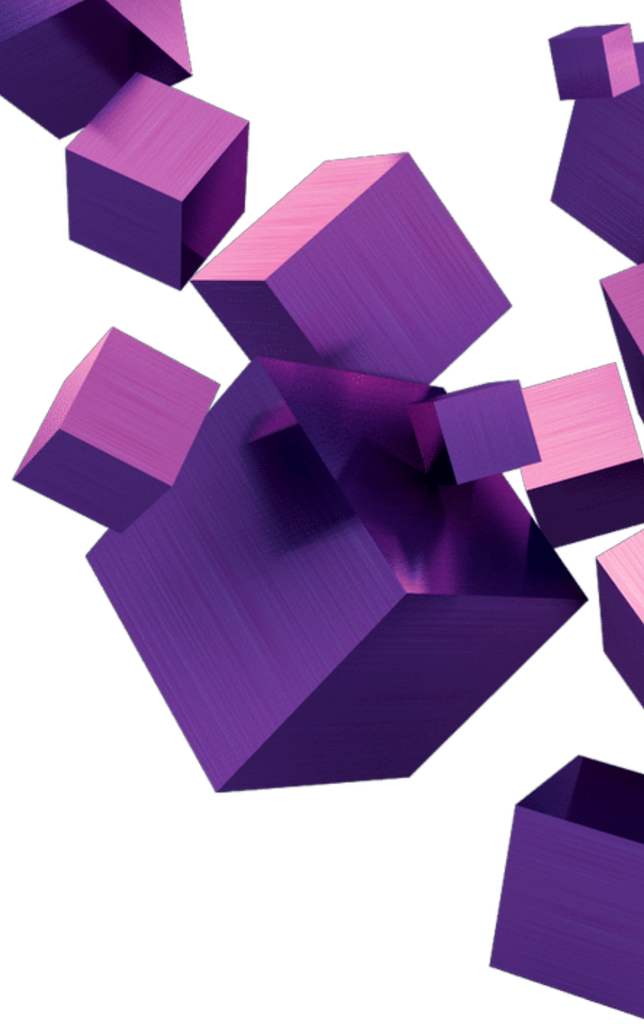
	checking_status	credit_history	purpose	savings_status	employment	personal_status	other_parties	property_magnitude	other_payment_plans	housing	foreign_worker	duration	credit_amount	age	class
0	1	1	6	4	3	3	2	3	1	1	1	6.0	1169.0	67.0	1
1	0	3	6	2	0	0	2	3	1	1	1	48.0	5951.0	22.0	0
2	3	1	2	2	1	3	2	3	1	1	1	12.0	2096.0	49.0	1
3	1	3	3	2	1	3	1	1	1	0	1	42.0	7882.0	45.0	1
4	1	2	4	2	0	3	2	2	1	0	1	24.0	4870.0	53.0	0

- Removing outliers above positive threshold 3 in numerical column using zscore
- Apply Standardization on Numerical Features

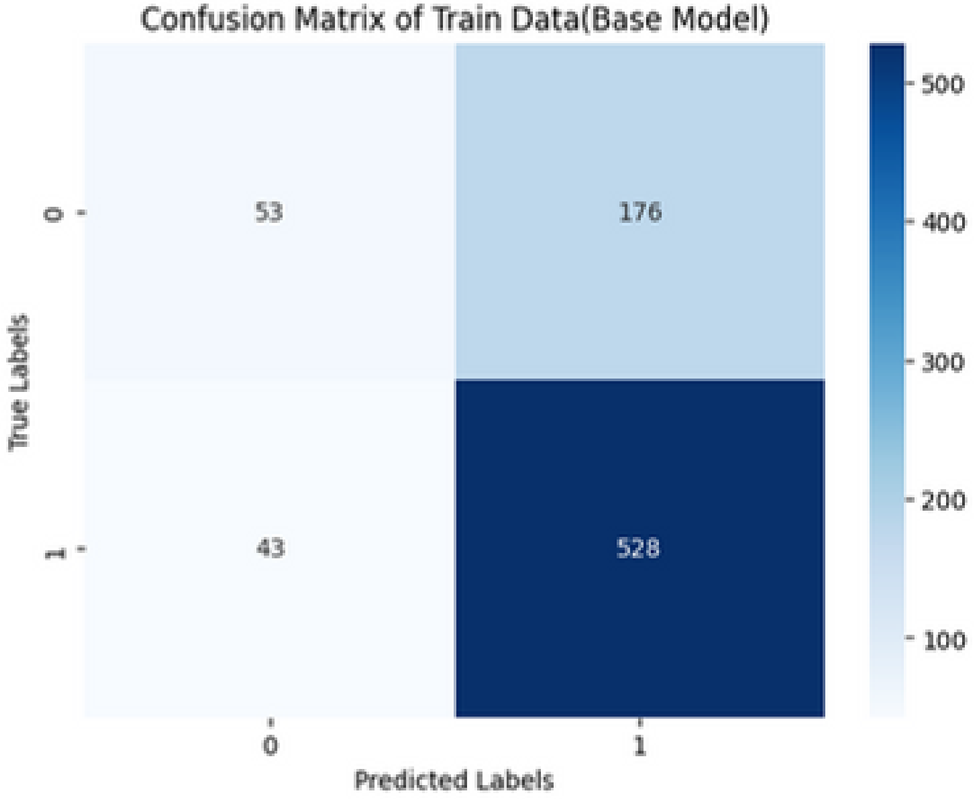
	checking_status	credit_history	purpose	savings_status	employment	personal_status	other_parties	property_magnitude	other_payment_plans	housing	foreign_worker	duration	credit_amount	age	class
0	1	1	6	4	3	3	2	3	1	1	1	-1.236478	-0.745131	2.766456	1
1	0	3	6	2	0	0	2	3	1	1	1	2.248194	0.949817	-1.191404	0
2	3	1	2	2	1	3	2	3	1	1	1	-0.738668	-0.416562	1.183312	1
3	1	3	3	2	1	3	1	1	1	0	1	1.750384	1.634247	0.831502	1
4	1	2	4	2	0	3	2	2	1	0	1	0.256953	0.566664	1.535122	0

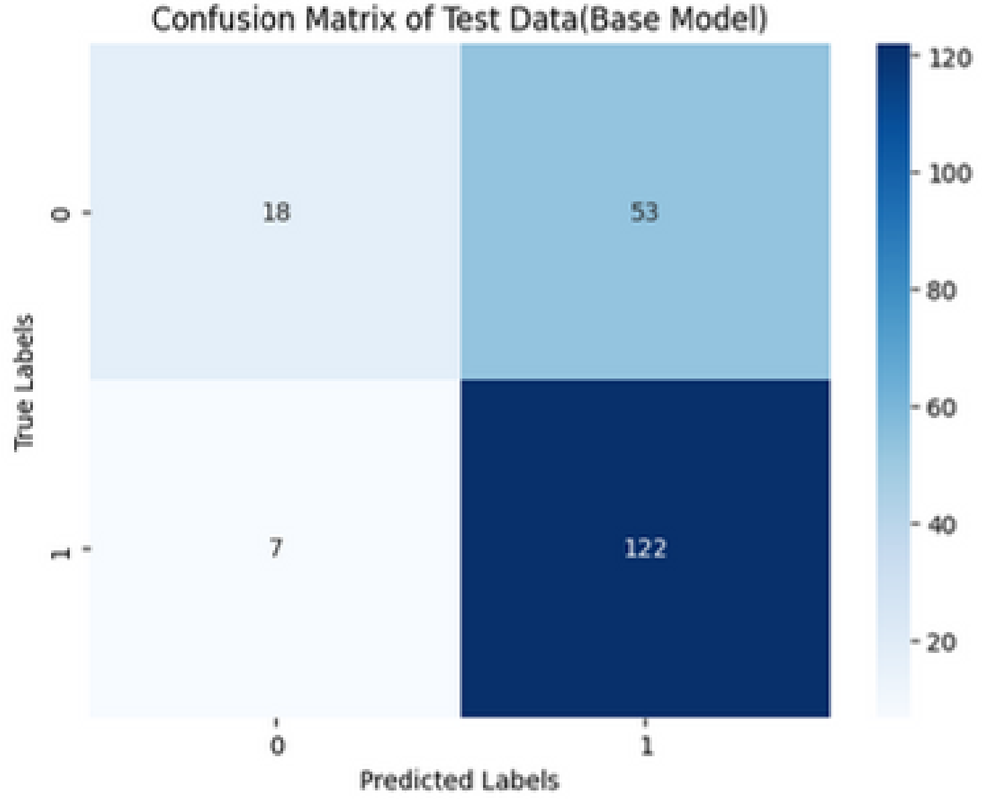
Models Outcome





Base Model (Logistic Regression)





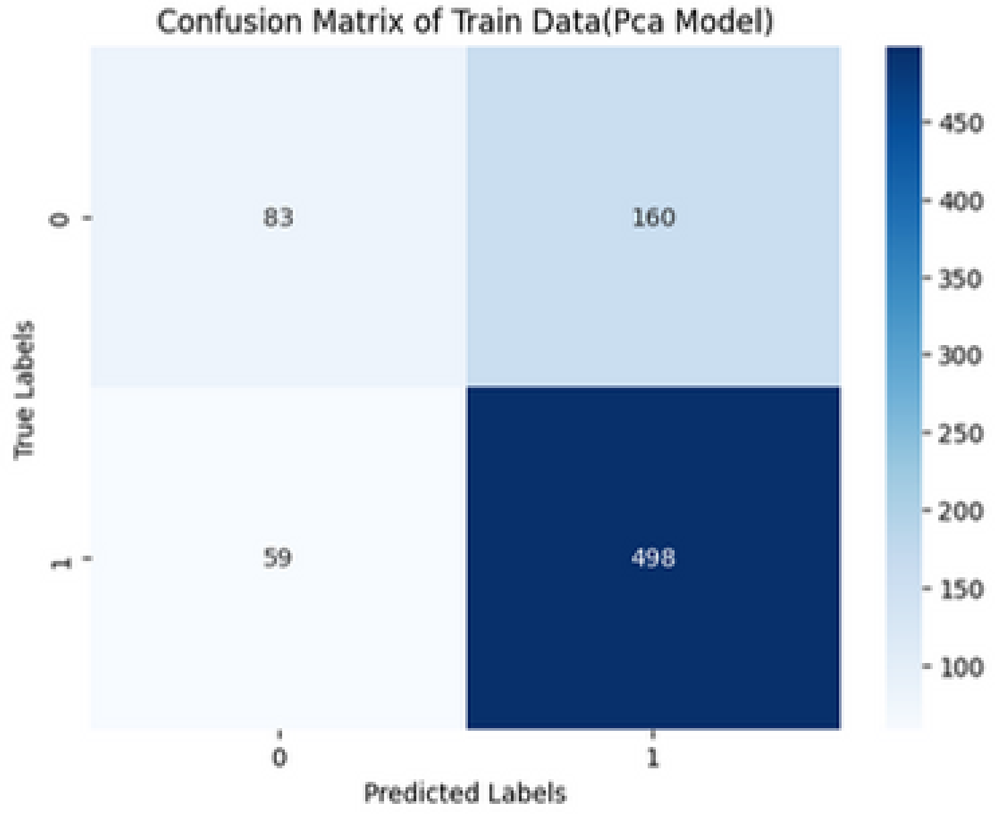
Classification Report of Train Data (Base Model):

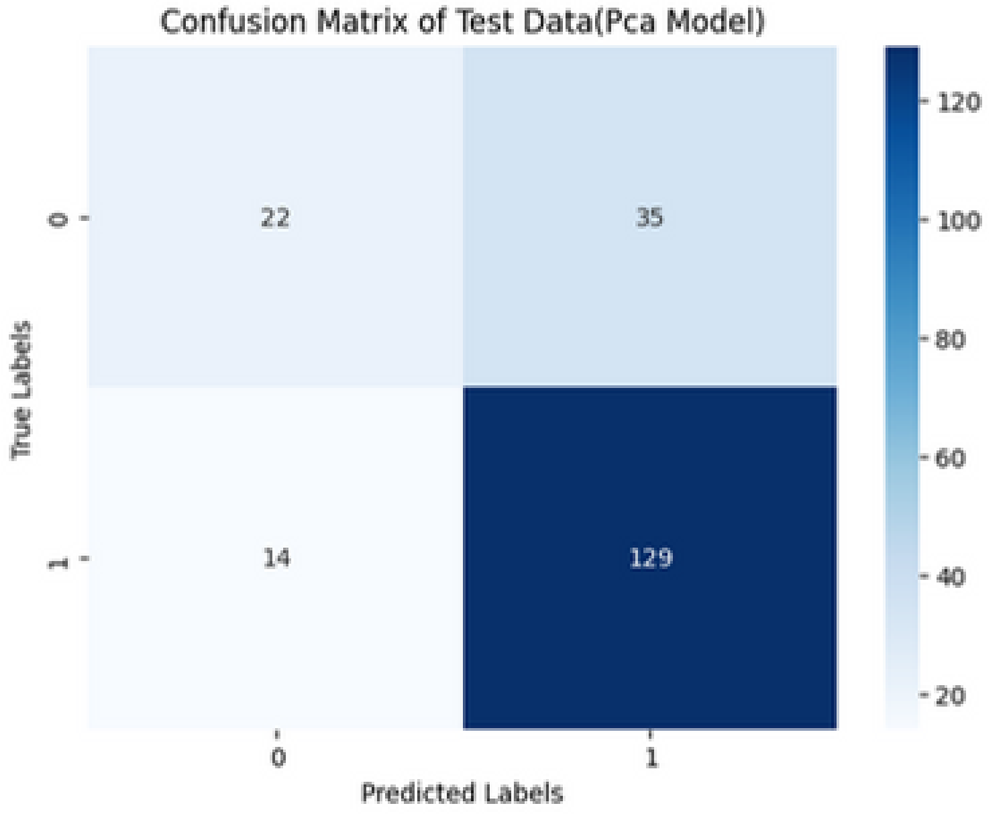
	precision	recall	f1-score	support
0	0.55	0.23	0.33	229
1	0.75	0.92	0.83	571
accuracy			0.73	800
macro avg	0.65	0.58	0.58	800
weighted avg	0.69	0.73	0.68	800

Classification Report of Test Data (Base Model):

	precision	recall	f1-score	support
0	0.72	0.25	0.38	71
1	0.70	0.95	0.80	129
accuracy			0.70	200
macro avg	0.71	0.60	0.59	200
weighted avg	0.71	0.70	0.65	200

PCA Then BaseLineModel





Classification Report of Train Data (Pca Model):

	precision	recall	f1-score	support
0	0.26	0.16	0.20	235
1	0.70	0.81	0.75	565
accuracy			0.62	800
macro avg	0.48	0.49	0.47	800
weighted avg	0.57	0.62	0.59	800

Classification Report of Test Data (Pca Model):

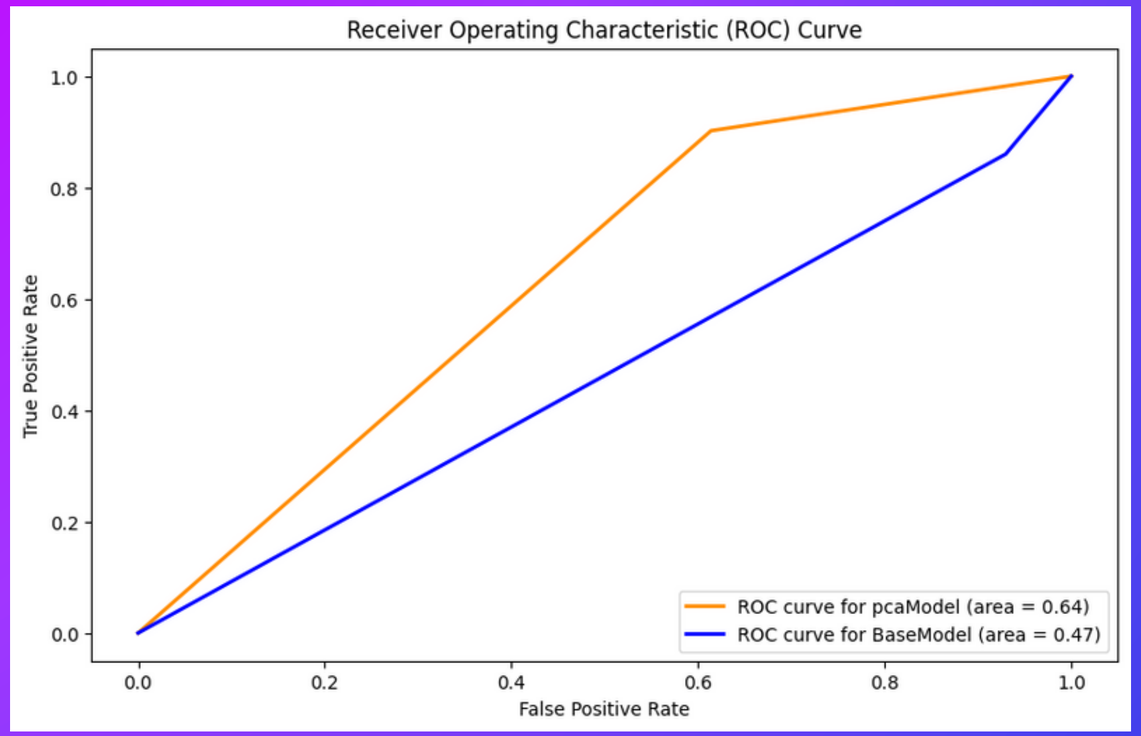
	precision	recall	f1-score	support
0	0.22	0.12	0.16	65
1	0.65	0.79	0.72	135
accuracy			0.57	200
macro avg	0.44	0.46	0.44	200
weighted avg	0.51	0.57	0.53	200

Compare Between the models

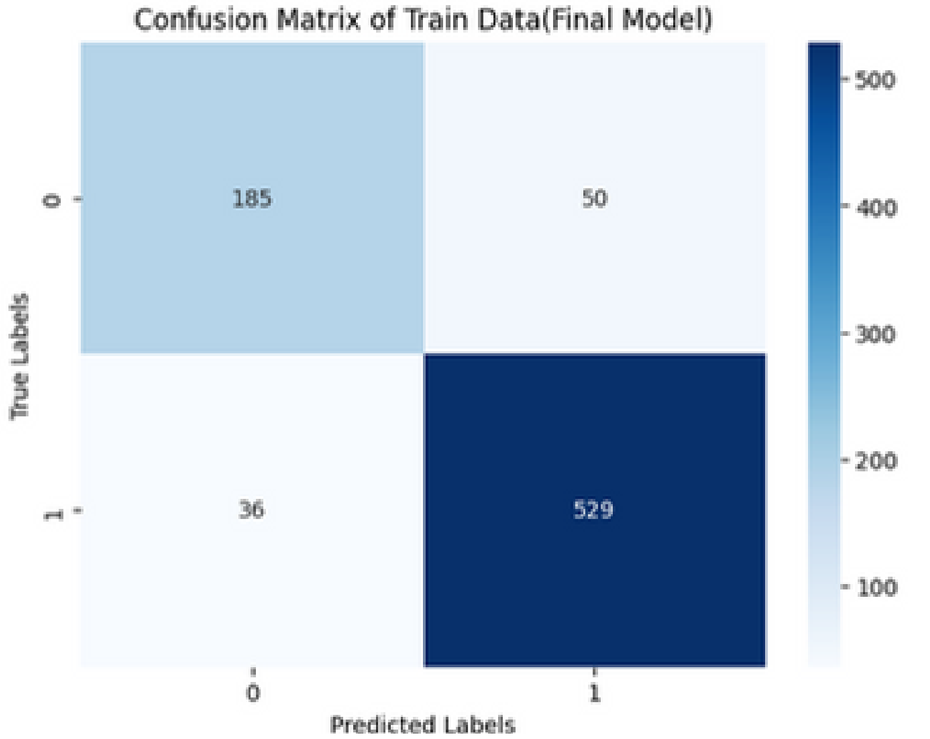
Cost Matrix Score

	Dataset	CostScore(BaseModel)	CostScore(pcaModel)
0	Train	904	859
1	Test	234	189

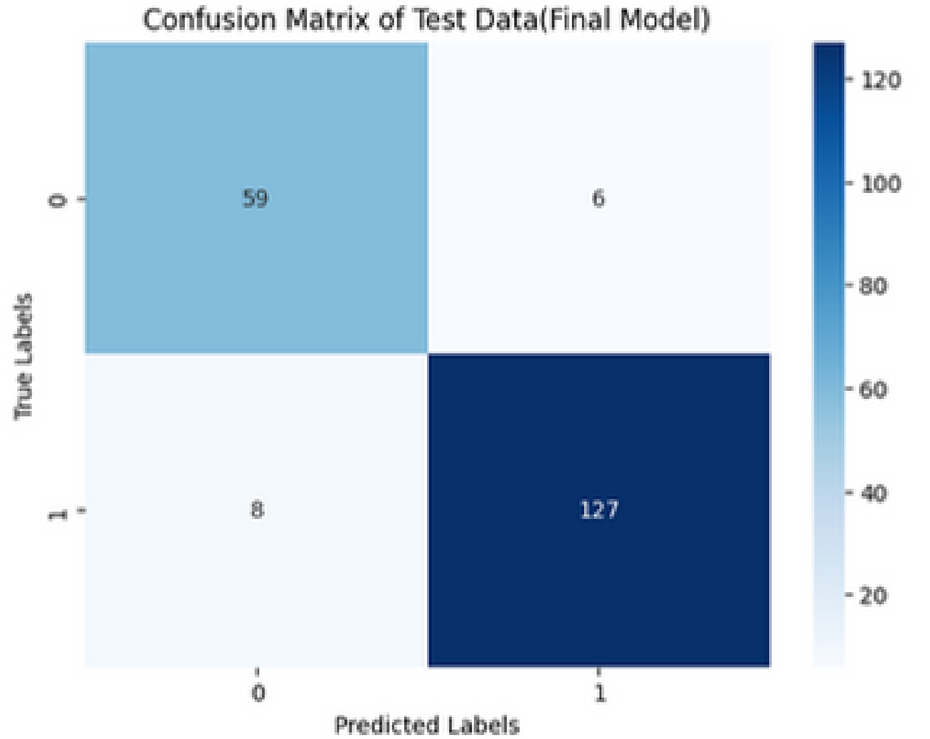
Area Under the curve



AutoEncoder with latent features 5 Then BaseLineModel



	precision	recall	f1-score	support
0	0.84	0.79	0.81	235
1	0.91	0.94	0.92	565
accuracy			0.89	800
macro avg	0.88	0.86	0.87	800
weighted avg	0.89	0.89	0.89	800

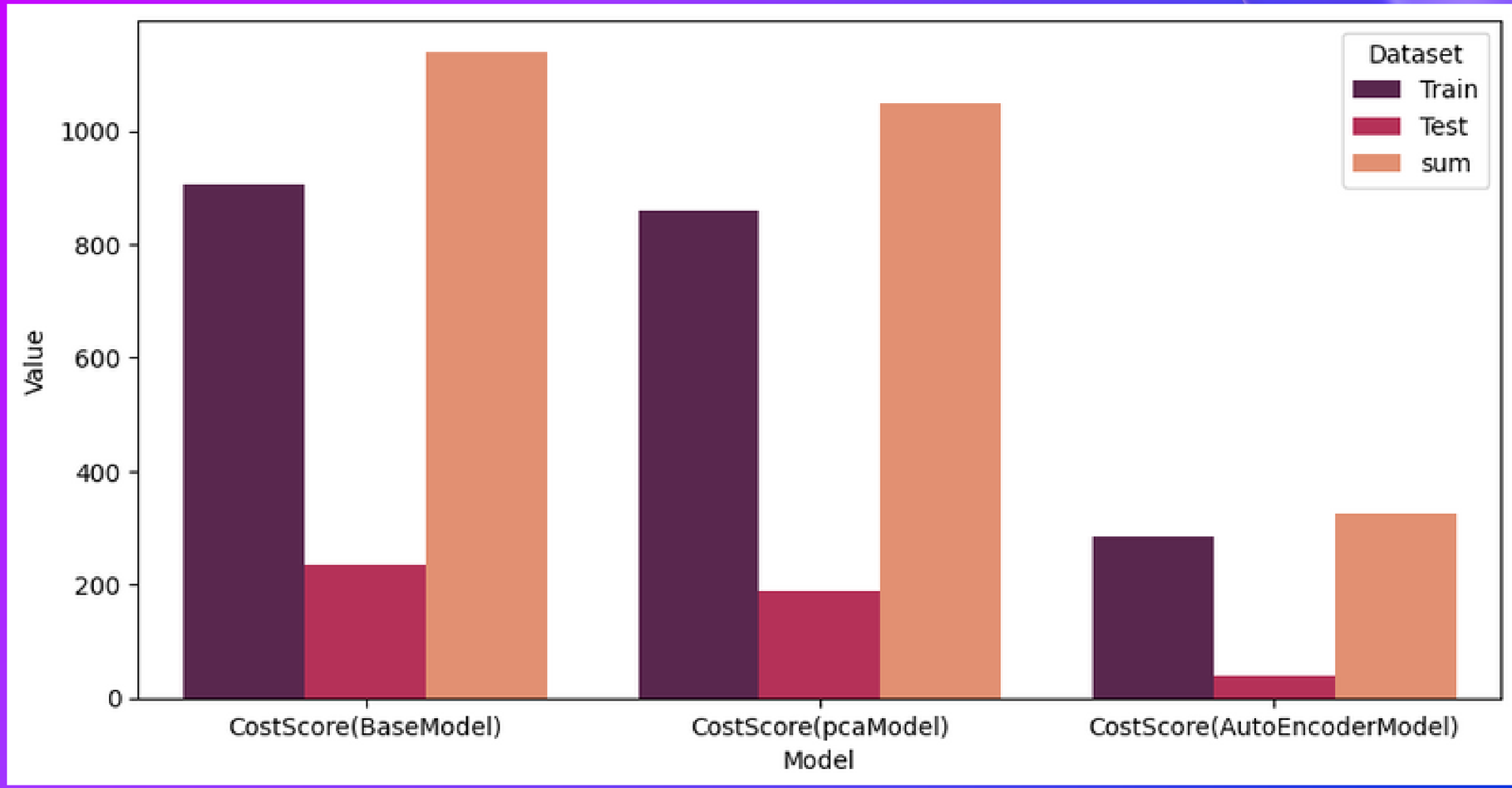


	precision	recall	f1-score	support
0	0.88	0.91	0.89	65
1	0.95	0.94	0.95	135
accuracy			0.93	200
macro avg	0.92	0.92	0.92	200
weighted avg	0.93	0.93	0.93	200

Compare Between the models

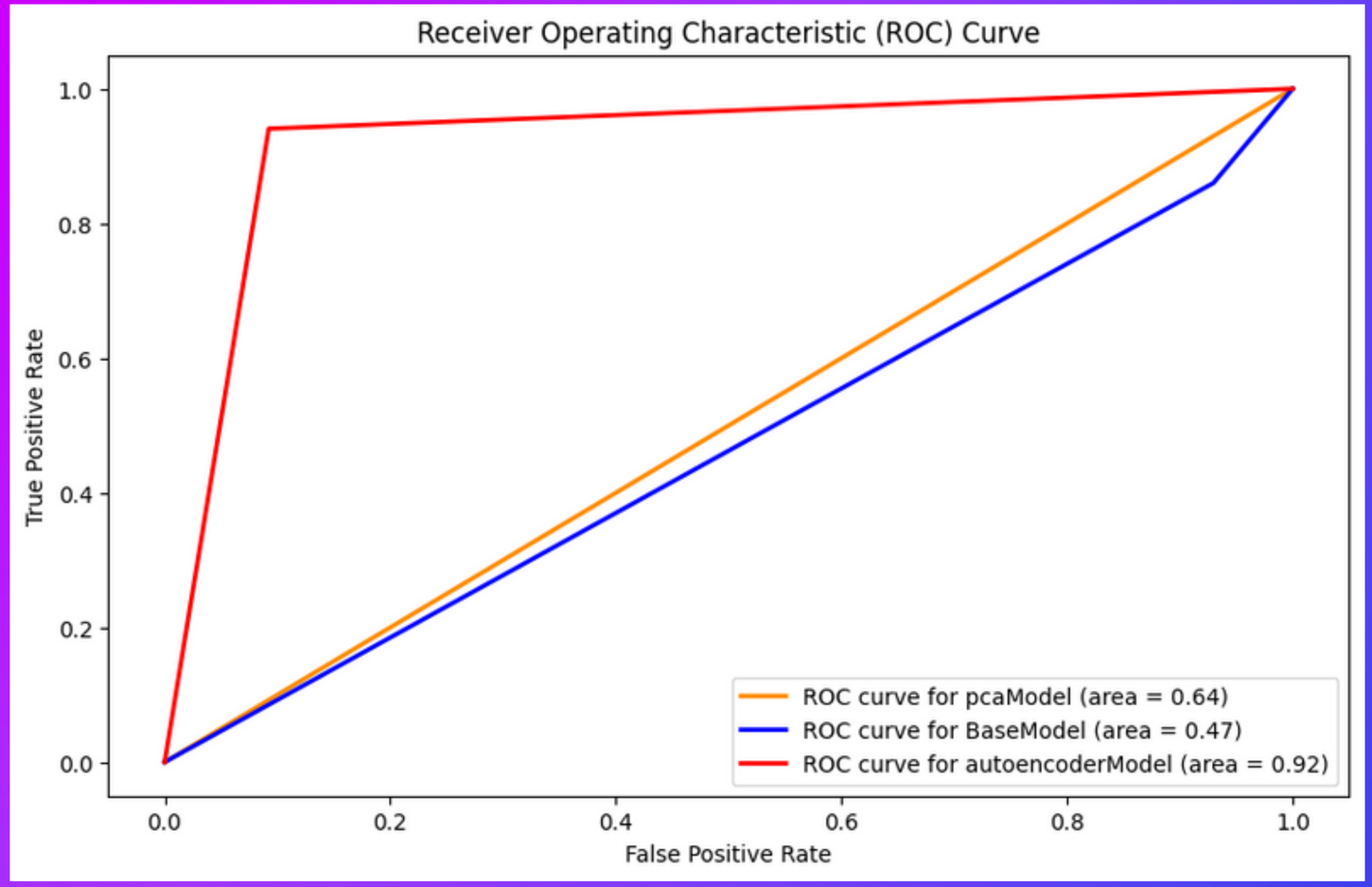
Cost Matrix Score

	Model	Train	Test	sum
0	CostScore(BaseModel)	904	234	1138
1	CostScore(pcaModel)	859	189	1048
2	CostScore(AutoEncoderModel)	286	38	324



Compare Between the models

Area Under the curve



Recommendations and Next steps

01

I suggest expanding the dataset size and addressing the class imbalance,to Improved Model Generalization

02

I think by using autoencoder to compress the features it’s the best way for that.

03

I think for the management part to be sure that user financial status will be good compared to amount of credit

04

Data points that important is salary and any feature indicate how this person lifestyle or how much needs per month.

Thank you