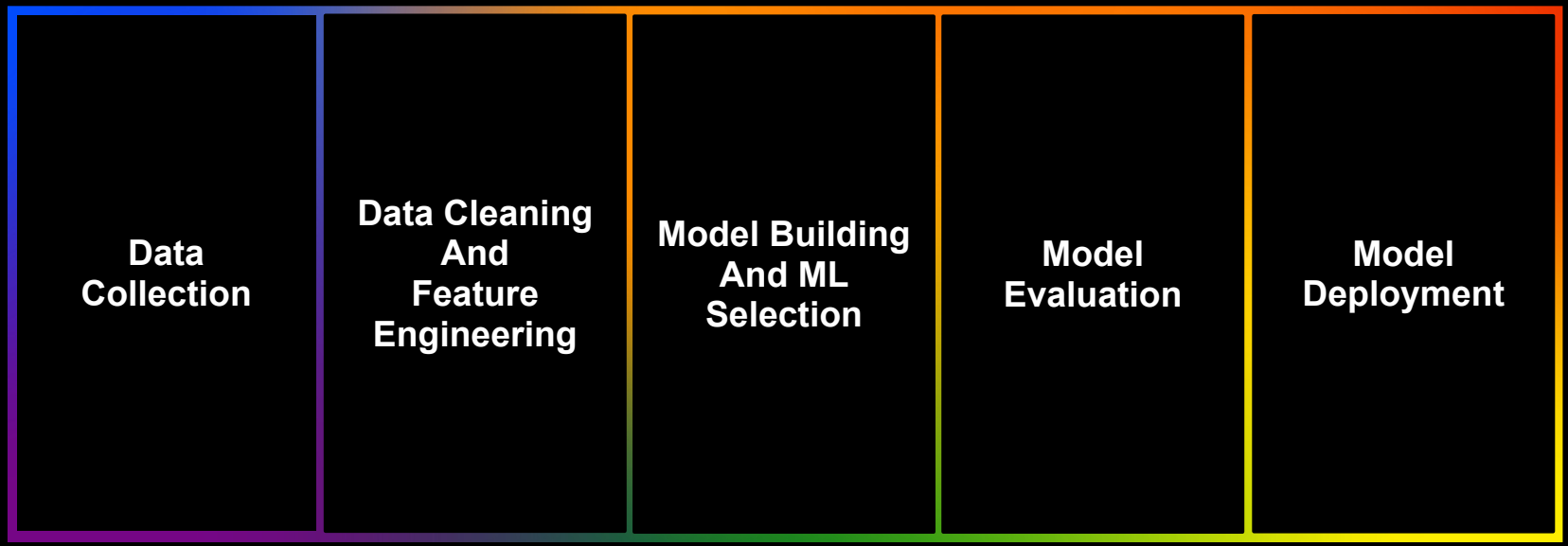
Predicting Customer Transactions

An Analysis Using Gaussian Naive Bayes Classification

Maxen Niedermier | Deep Dive Data Science Cohort 13 | 2.15.2024


Data Collection
Data Cleaning And Feature Engineering
Model Building And ML Selection
Model Evaluation
Model Deployment

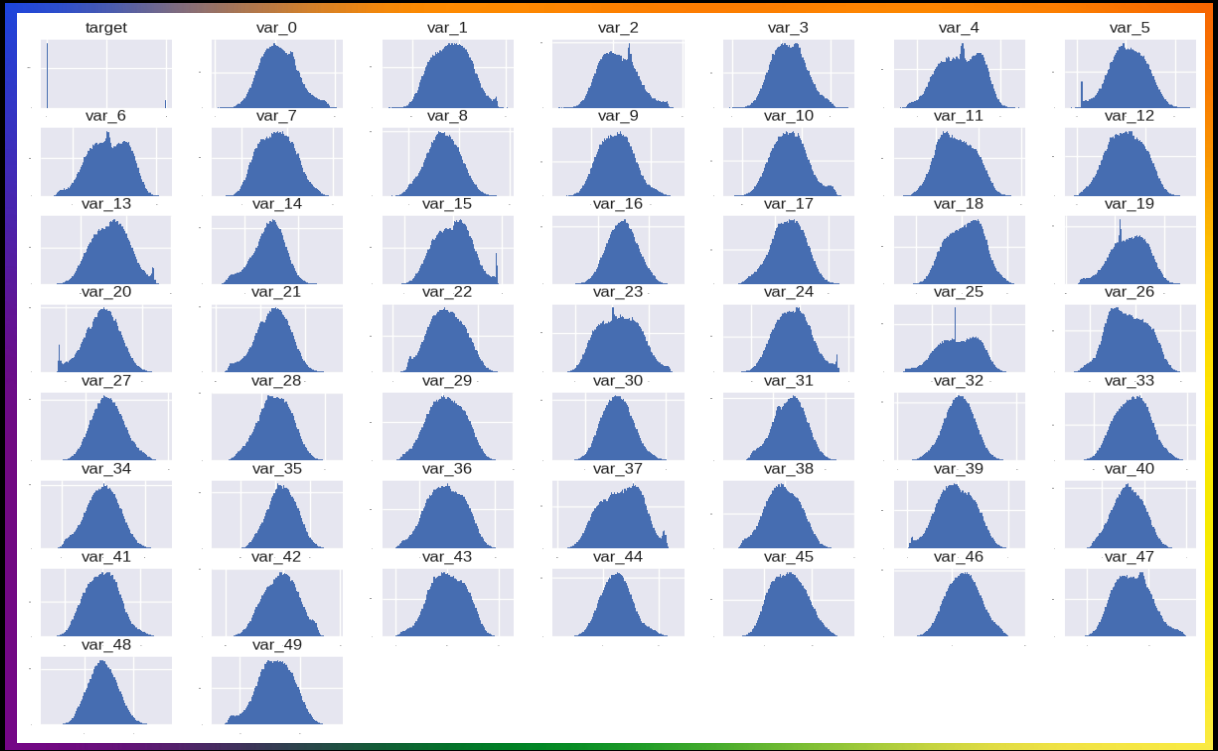
Data Types

180000 Rows and 53 Columns

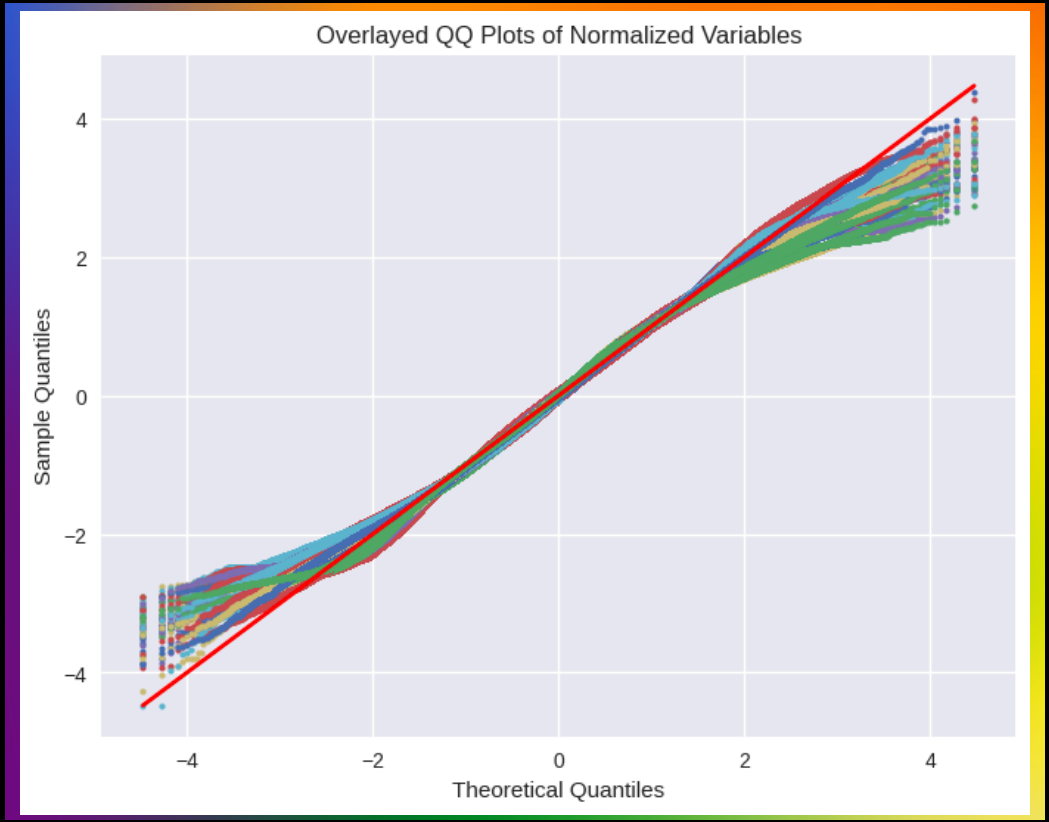
Type: float64	Type: int64	Type: object
var_0, …,var_49	Unnamed: 0, target	ColumnID_Code
Column Count: 50	Column Count: 2	Column Count: 1

	Unnamed: 0	ID_code	target	var_0	var_1	var_2	var_3	var_4
0	0	train_0	0	8.9255	-6.7863	11.9081	5.1187	5.7470
1	1	train_1	0	11.5006	-4.1473	13.8588	5.6208	8.0851

Data Distribution of variables

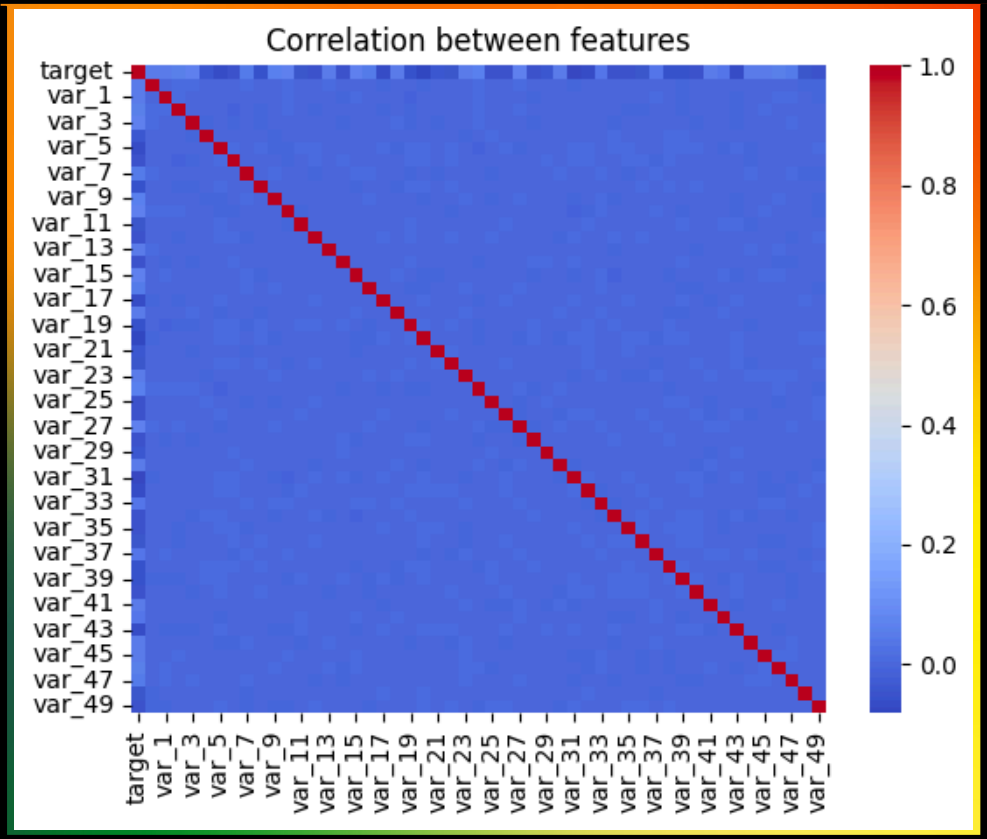


Data Distribution of variables



Correlation

Assumes that there is no statistical correlation between any features



Model Building

The Data

Feature Set 1

Target Values
1: 18040
0: 161960

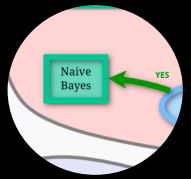
Biased data
90 to 10

Feature Set 2

Target Values
1: 18040
0: 18040

Resampled Balanced data
1 to 1

Model Selection



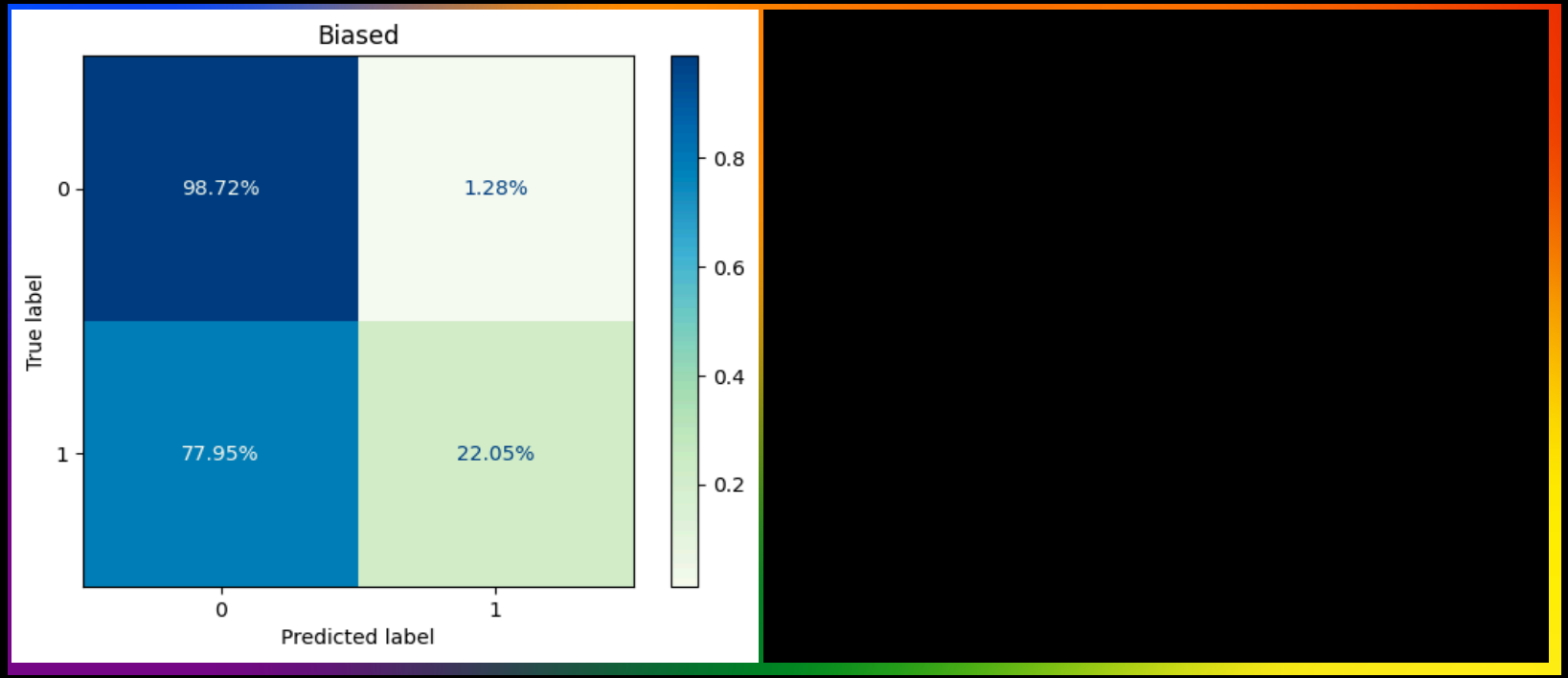
With Independent Variables
That are normally distributed

Classification

Unsupervised

Naive Bayes

Model Building



The number of times the model predicted the real target:
-- 40944 out of 45000 points. –
91% Accuracy

The number of times the model predicted the real target:
-- 5469 out of 7216 points --
75% Accuracy

Conclusion and Future Work

Balancing reduces error

Conclusion and Future Work

Experiment with more models

Raise accuracy
without increasing error

Conclusion and Future Work

Addressing bias

Resampling
Weighted Loss Functions
Anomaly Detection Techniques
SMOTE (Synthetic Minority Over-sampling Technique)

Follow for more

GITHUB

Maxenlee/DeepDive

LinkedIn

maxenlee