Predicting default of credit card clients

Roger Bukuru

2024-03-11

Contents

1 Import Data **2**

2 Introduction **2**

3 Exploratory Data Analysis **2**

3.1 Feature Analysis 2

3.2 Data Cleaning 6

3.3 Feature Selection 10

3.4 Feature Extraction 10

4 Model Engineering **11**

5 Model Evaluation **11**

5.1 Conclusion 11

1 Import Data

2 Introduction

One of the many functions of a bank is the ability to provide its customers with loans or credit facilities. These channels provide banks with the opportunity to generate additional revenues by means of gaining interest off the back of these financial instruments.

However these instruments have an intrinsic risk once awarded to a customer, customers may default in their repayments of such loans or credit facilities. In this research, we will focus on credit card facilities, specifically credit card clients in Taiwan. We will be exploring the probability of clients defaulting on their credit card facility given a set of features such as their demographic information (like gender, education, marital status, and age) to financial behaviors (including payment history, bill statement amounts, and previous payment amounts). The exploration will aim to provide statistical techniques to better manage credit risk default. Credit risk here means, the probability of a delay in the repayment of the credit granted.

Furthermore, we aim to explore given the above-mentioned features, which features/factors are more likely to affect a customer defaulting e.g. are younger customers more likely to default than older clients, and given these factors what is the probability of a customer defaulting? This will be done by exploring established techniques such as employing customer risk segmentation by conducting clustering techniques such as k-means and Linear Discriminant Analysis (LDA) techniques[**?**]. To assess the effectiveness of these techniques we will explore and compare them to Support Vector Machines (SVMs). As our data has many input variables, feature extraction methods such as PCA and Autoencoders will be explored before performing the above clustering and classification techniques. The repository of this project is available at

3 Exploratory Data Analysis

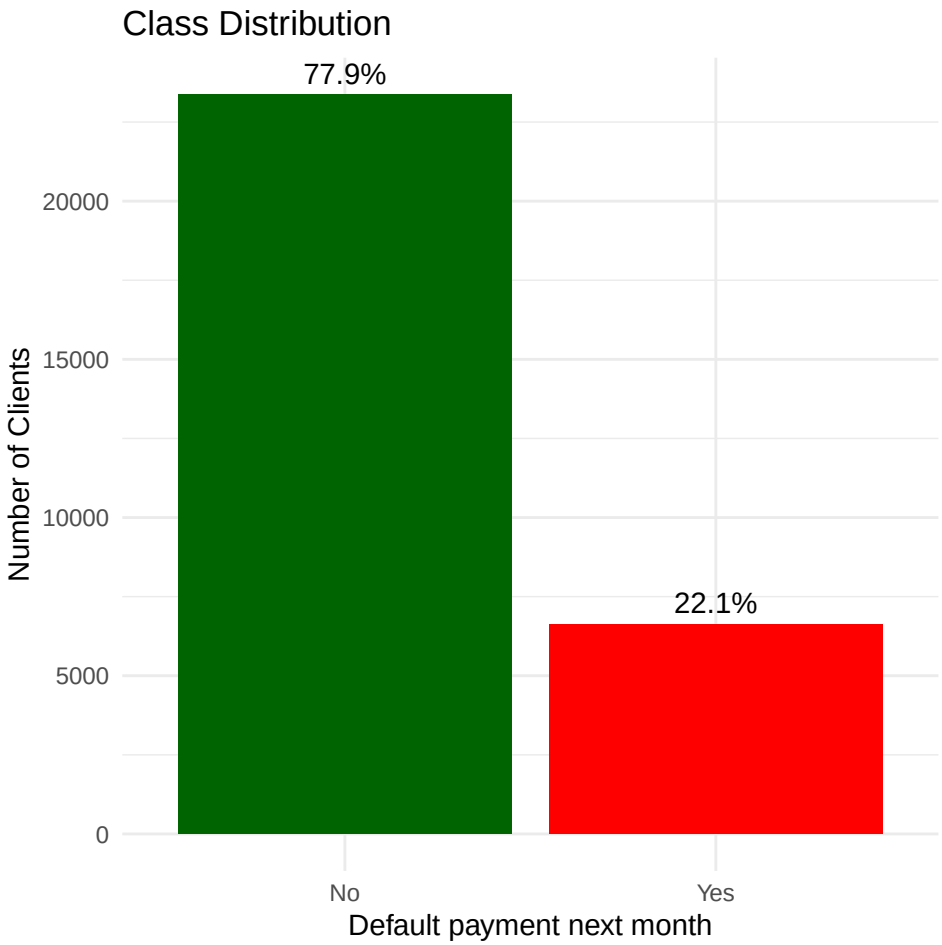
3.1 Feature Analysis

```
##  [1] "ID"                        "LIMIT_BAL"
##  [3] "SEX"                       "EDUCATION"
##  [5] "MARRIAGE"                  "AGE"
##  [7] "PAY_0"                     "PAY_2"
##  [9] "PAY_3"                     "PAY_4"
## [11] "PAY_5"                     "PAY_6"
## [13] "BILL_AMT1"                 "BILL_AMT2"
## [15] "BILL_AMT3"                 "BILL_AMT4"
```

```
## [17] "BILL_AMT5"                    "BILL_AMT6"
## [19] "PAY_AMT1"                     "PAY_AMT2"
## [21] "PAY_AMT3"                     "PAY_AMT4"
## [23] "PAY_AMT5"                     "PAY_AMT6"
## [25] "default payment next month"
```

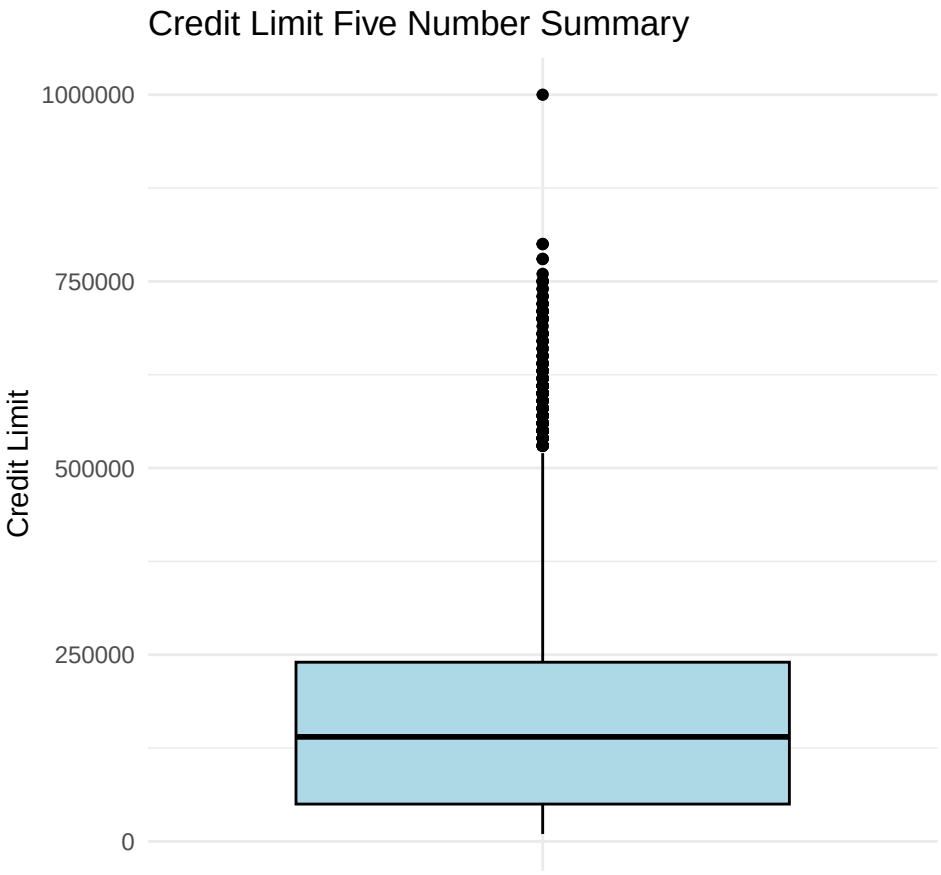
	vars	n	mean	sd	min	max	range	se
ID	1	30000	15000.50	8660.40	1	30000	29999	50.00
LIMIT_BAL	2	30000	167484.32	129747.66	10000	1000000	990000	749.10
SEX	3	30000	1.60	0.49	1	2	1	0.00
EDUCATION	4	30000	1.85	0.79	0	6	6	0.00
MARRIAGE	5	30000	1.55	0.52	0	3	3	0.00
AGE	6	30000	35.49	9.22	21	79	58	0.05
PAY_0	7	30000	-0.02	1.12	-2	8	10	0.01
PAY_2	8	30000	-0.13	1.20	-2	8	10	0.01
PAY_3	9	30000	-0.17	1.20	-2	8	10	0.01
PAY_4	10	30000	-0.22	1.17	-2	8	10	0.01
PAY_5	11	30000	-0.27	1.13	-2	8	10	0.01
PAY_6	12	30000	-0.29	1.15	-2	8	10	0.01
BILL_AMT1	13	30000	51223.33	73635.86	-165580	964511	1130091	425.14
BILL_AMT2	14	30000	49179.08	71173.77	-69777	983931	1053708	410.92
BILL_AMT3	15	30000	47013.15	69349.39	-157264	1664089	1821353	400.39
BILL_AMT4	16	30000	43262.95	64332.86	-170000	891586	1061586	371.43
BILL_AMT5	17	30000	40311.40	60797.16	-81334	927171	1008505	351.01
BILL_AMT6	18	30000	38871.76	59554.11	-339603	961664	1301267	343.84
PAY_AMT1	19	30000	5663.58	16563.28	0	873552	873552	95.63
PAY_AMT2	20	30000	5921.16	23040.87	0	1684259	1684259	133.03
PAY_AMT3	21	30000	5225.68	17606.96	0	896040	896040	101.65
PAY_AMT4	22	30000	4826.08	15666.16	0	621000	621000	90.45
PAY_AMT5	23	30000	4799.39	15278.31	0	426529	426529	88.21
PAY_AMT6	24	30000	5215.50	17777.47	0	528666	528666	102.64
default payment next month	25	30000	0.22	0.42	0	1	1	0.00

- Education and Marriage have some unreported categories


Class Distribution
77.9%
22.1%
20000
15000
10000
5000
0
Number of Clients
No
Yes
Default payment next month

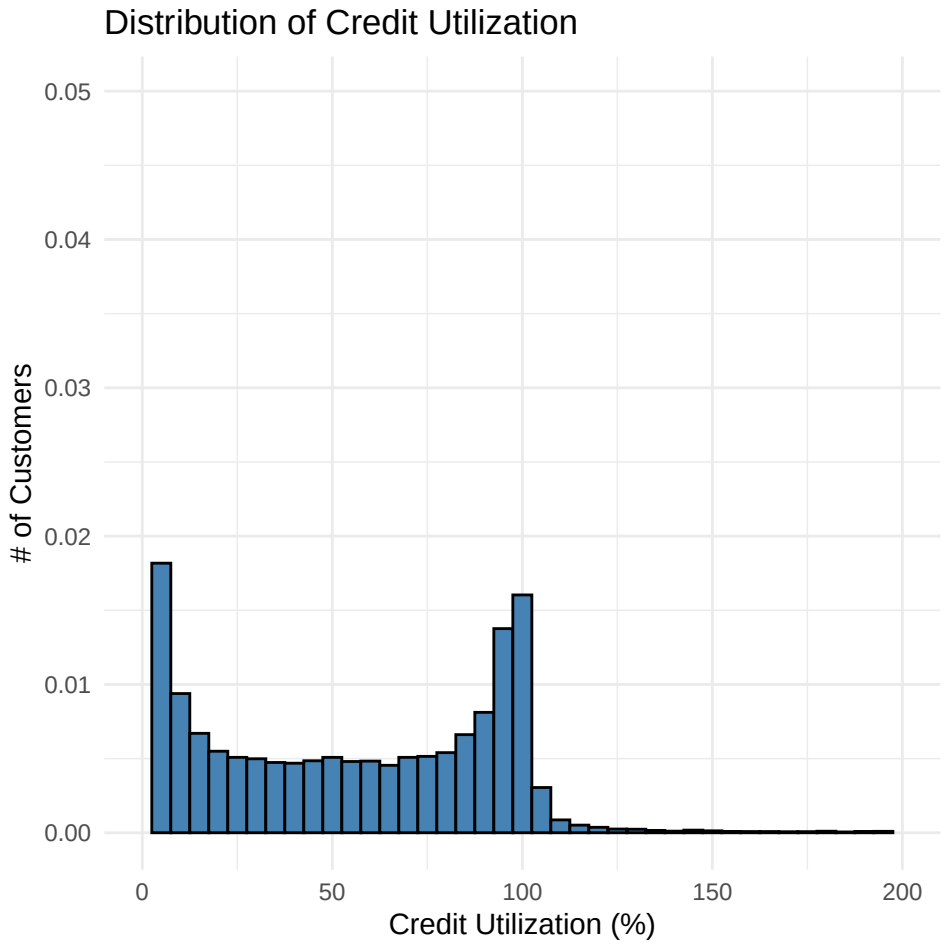
-

```
## [1] 1e+04 1e+06
```



- The minimum credit limit balance was NT$ 10 000
- 25% of customers had credit limit balances of NT$ 50 000 or less, indicating that 75% of clients had credit limit balances above NT$ 50,000.
- The median credit limit was NT$ 140 000, indicating that 50% of customers have credit limit balances less than NT$ 140 000 whilst another 50% of customers have credit limit balances above NT$ 140 000.
- 75% of customers had credit limit balances below NT$ 240 000, whilst 25% of customers had credit limit balances above NT$ 240 000
- The maximum credit limit amount was NT$ 1 000 000
- The average male credit limit balance was NT$ 163519.82 and the average female credit limit balance was NT$ 170086.46
- Customer ages ranged from 21 to 79 years old, with the average age being 35 years old.

The summary above suggests the following; there is significant variability in the credit limit balances that are given to customers as the limits range from NT$ 10 000 to NT$ 1 000 000. Furthermore, we note that the median value is closer to the third quantile than it is to the first, indicating that the data is right-skewed, which means that there is a large number of customers who have credit limits on the lower end of the range, with fewer individuals having higher credit limits. The significant difference between the maximum value and the third quantile finally indicates that outliers exist on the higher end of the credit limit balances.



- There are a few customers that have exceeded their credit limit facility. Indicating several default risk customers.
- Overall the distribution is left-skewed, indicating that more customers have utilized a large portion of their credit facility than less. This could indicate that several customers could be at risk of defaulting as they owe more money.

3.2 Data Cleaning

MARRIAGE	Total
0	54
1	13659
2	15964
3	323

EDUCATION	Total
0	14
1	10585
2	14030
3	4917
4	123
5	280
6	51

MARRIAGE	Total
1	13659
2	15964
3	377

EDUCATION	Total
1	10585
2	14030
3	4917
4	468

```
# Payment History where value is -2 and 0, aggregate under -1

payment_history = c("PAY_1", "PAY_2", "PAY_3", "PAY_4", "PAY_5","PAY_6")

for (p in payment_history) {
  indices = which(credit_default_data[,p] <= 0)
  credit_default_data[indices, p] = -1
}
```

```
payment_history_categorical = paste("PAY_", 1:6, sep = "")
categorical_feature_names = c("SEX", "EDUCATION", "MARRIAGE", "DEFAULT_PAYMENT_NEXT_MONT
credit_default_data = credit_default_data%>%
                      mutate(across(matches(categorical_feature_names), as.factor))
```

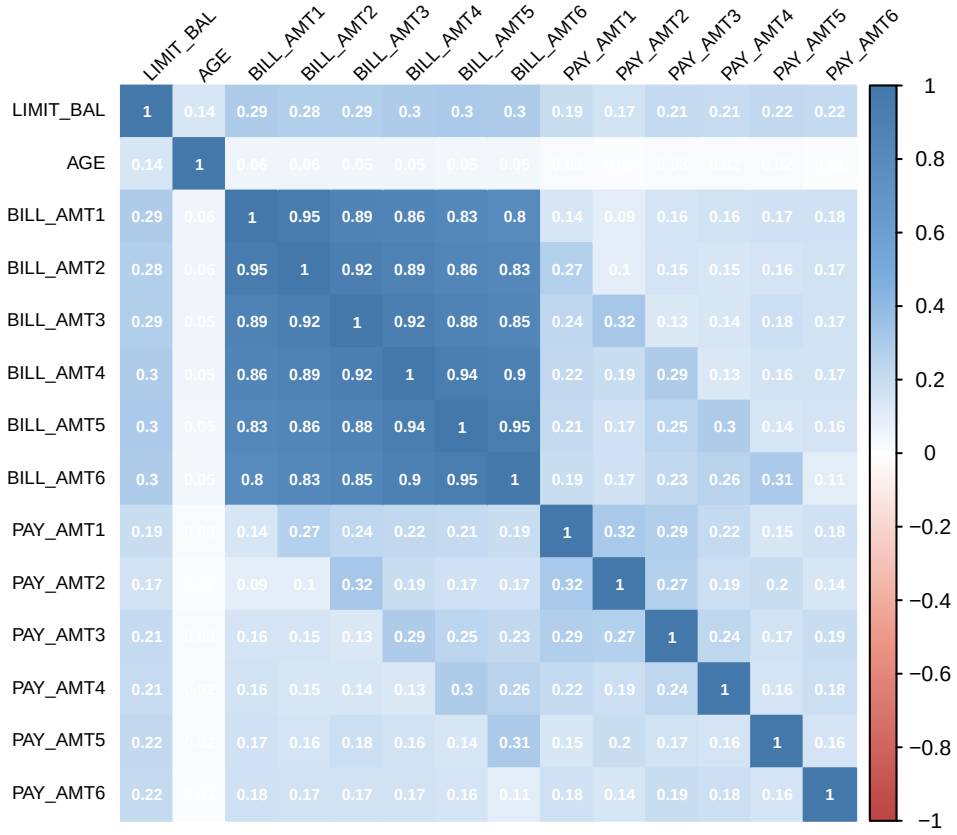
```r
#summary(credit_default_data)
# Split Data

set.seed(10032024)
dataSize = nrow(credit_default_data)
trainingSize = floor(0.75*dataSize)

trainingDataIndices = sample(seq_len(dataSize), size = trainingSize)

training_data = credit_default_data[trainingDataIndices,]
testing_data  = credit_default_data[-trainingDataIndices,]
```

```
##  SEX        EDUCATION MARRIAGE      PAY_1           PAY_2           PAY_3
##  1: 8903    1: 7946   1:10234   -1     :17404   -1     :19176   -1     :19355
##  2:13597    2:10553   2:12004   1      : 2725   2      : 2954   2      : 2835
##             3: 3663   3:  262   2      : 2019   3      :  231   3      :  195
##             4:  338             3      :  241   4      :   78   4      :   59
##                                 4      :   59   1      :   17   6      :   18
##                                 5      :   22   5      :   16   7      :   17
##                                 (Other):   30   (Other):   28   (Other):   21
##      PAY_4           PAY_5           PAY_6       DEFAULT_PAYMENT_NEXT_MONTH
##  -1     :19868   -1     :20264   -1     :20204   0:17537
##  2      : 2361   2      : 1971   2      : 2055   1: 4963
##  3      :  138   3      :  136   3      :  148
##  4      :   55   4      :   68   7      :   36
##  7      :   44   7      :   46   4      :   33
##  5      :   27   5      :   13   6      :   14
##  (Other):    7   (Other):    2   (Other):   10
```



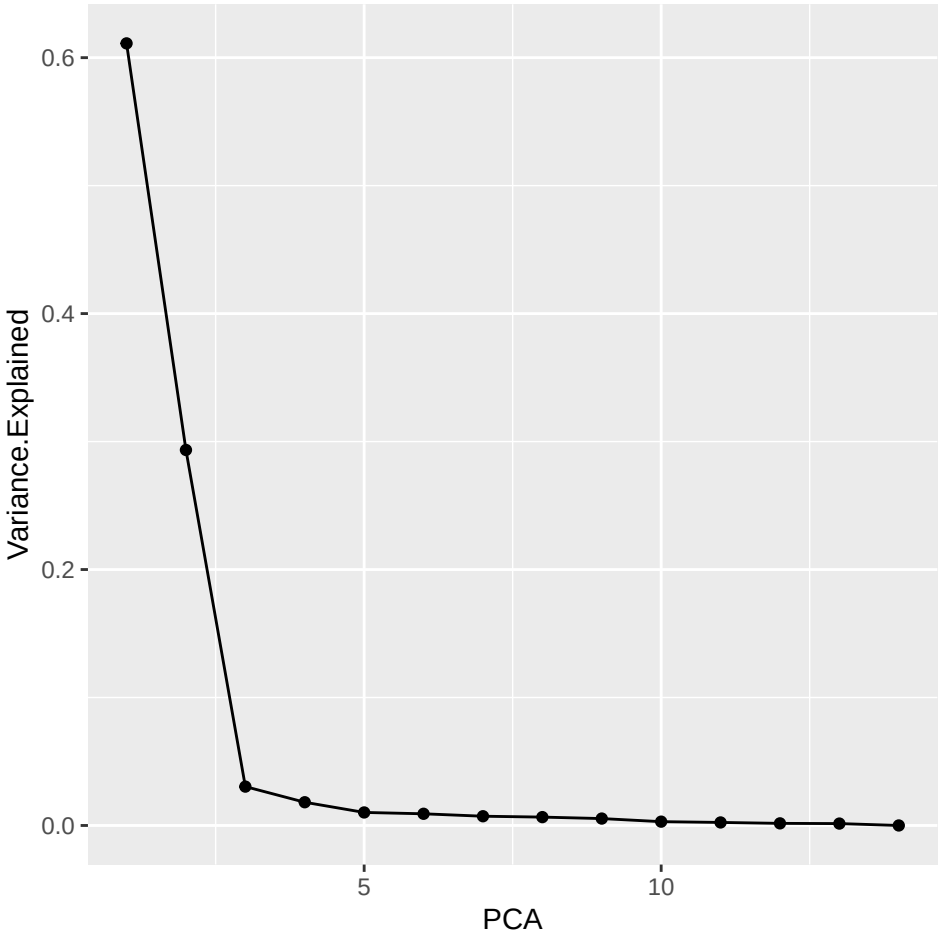
We observe from the correlation matrix, that some features have high correlations with each other

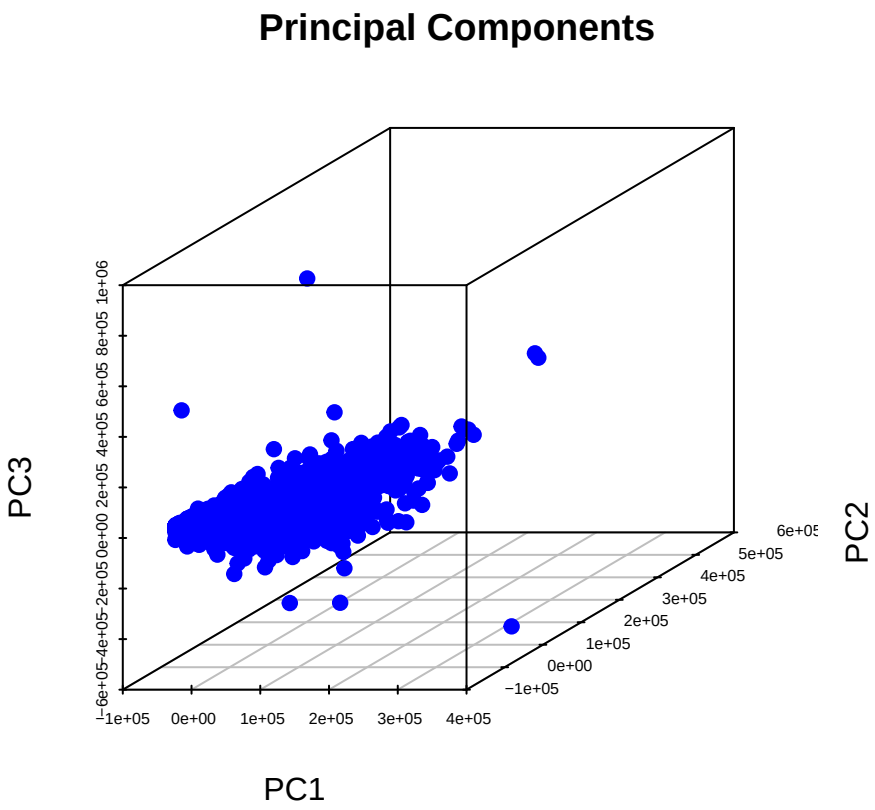
- BILL_AMT1 and BILL_AMT2 have p = 0.95
- BILL_AMT2 and BILL_AMT3 have p = 0.92
- BILL_AMT4 and BILL_AMT5 have p = 0.94

3.3 Feature Selection

3.4 Feature Extraction

3.4.1 Principal Component Analysis





Autoencoders

4 Model Engineering

5 Model Evaluation

5.1 Conclusion