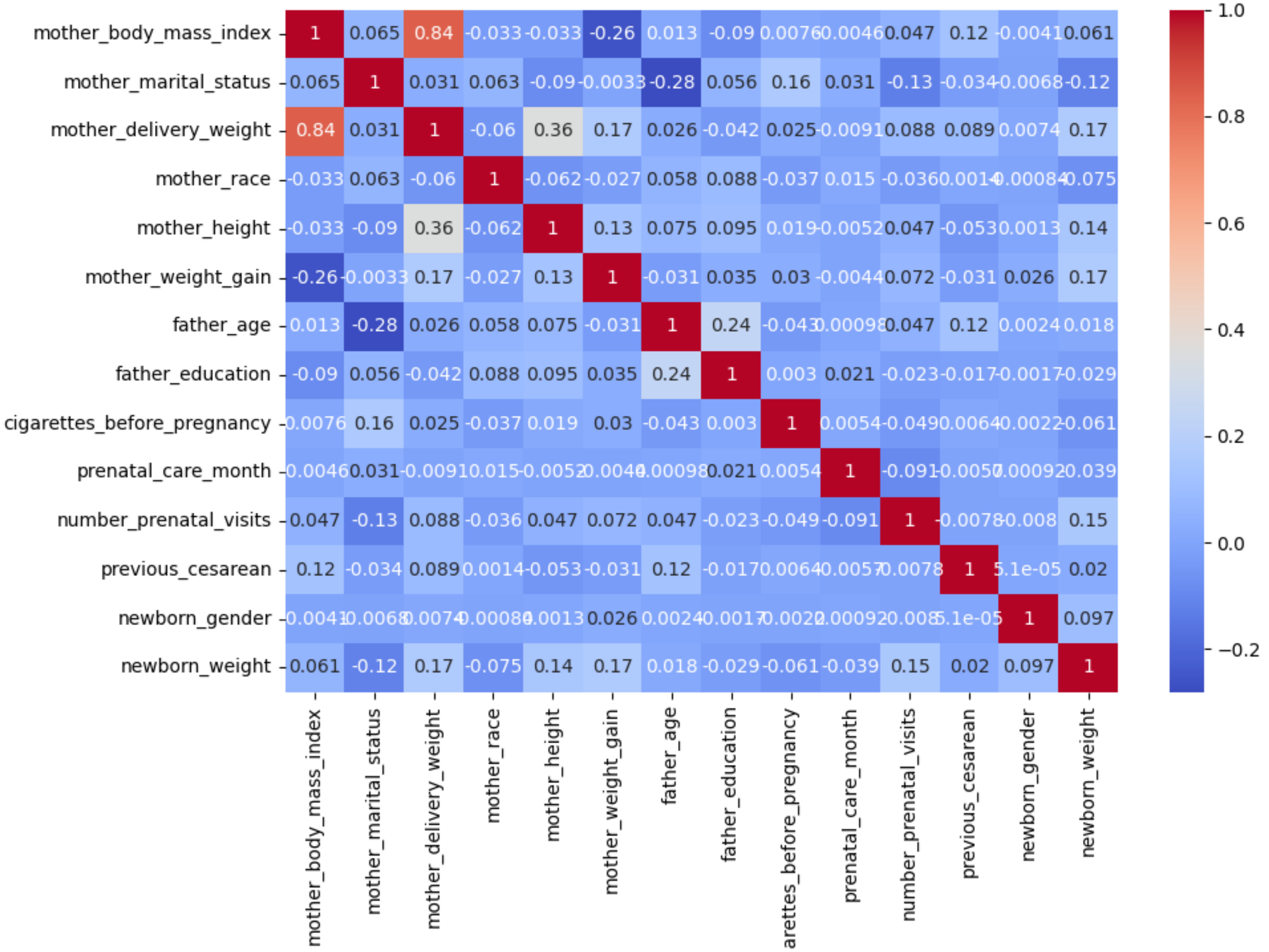
ML Project for Regression

Yiqing Hu 455858
Raffaele Ricci K-15305

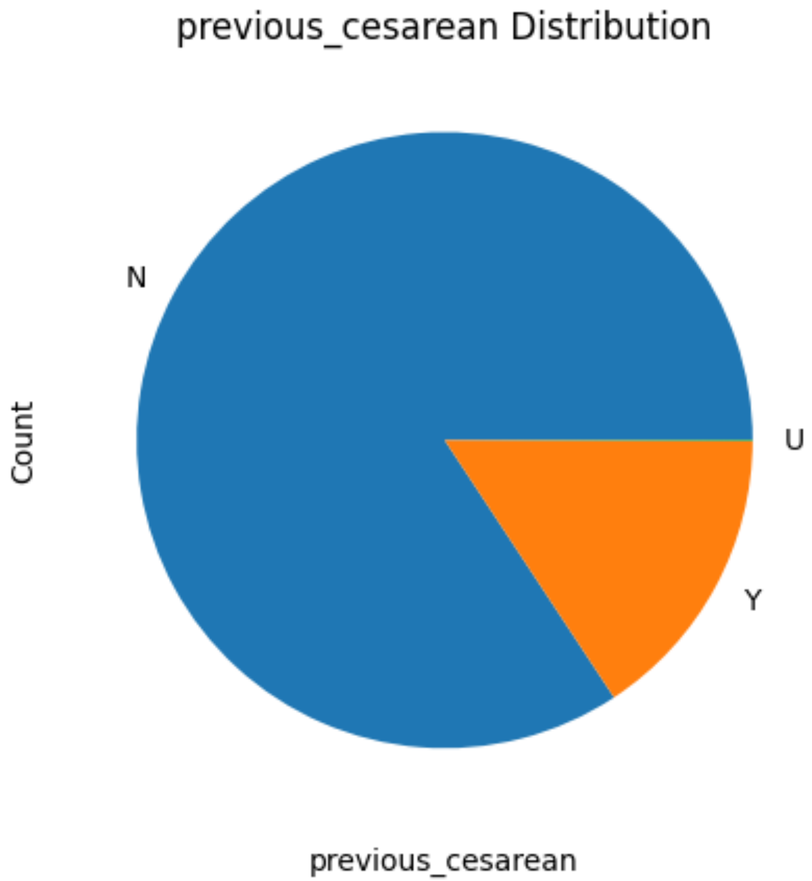
Data Checking

- Correlation
- Distribution
- NA


mother_body_mass_index 1 0.065 0.84 -0.033 -0.033 -0.26 0.013 -0.09 0.0076 0.0046 0.047 0.12 -0.0041 0.061
mother_marital_status 0.065 1 0.031 0.063 -0.09 -0.0033 -0.28 0.056 0.16 0.031 -0.13 -0.034 0.0068 -0.12
mother_delivery_weight 0.84 0.031 1 -0.06 0.36 0.17 0.026 -0.042 0.025 -0.0091 0.088 0.089 0.0074 0.17
mother_race -0.033 0.063 -0.06 1 -0.062 -0.027 0.058 0.088 -0.037 0.015 -0.036 0.0014 0.00084 0.075
mother_height -0.033 -0.09 0.36 -0.062 1 0.13 0.075 0.095 0.019 -0.0052 0.047 -0.053 0.0013 0.14
mother_weight_gain -0.26 -0.0033 0.17 -0.027 0.13 1 -0.031 0.035 0.03 -0.0044 0.072 -0.031 0.026 0.17
father_age 0.013 -0.28 0.026 0.058 0.075 -0.031 1 0.24 -0.043 0.00098 0.047 0.12 0.0024 0.018
father_education -0.09 0.056 -0.042 0.088 0.095 0.035 0.24 1 0.003 0.021 -0.023 -0.017 -0.0017 -0.029
cigarettes_before_pregnancy 0.0076 0.16 0.025 -0.037 0.019 0.03 -0.043 0.003 1 0.0054 -0.049 0.0064 0.0022 -0.061
prenatal_care_month 0.0046 0.031 -0.0091 0.015 -0.0052 0.0046 0.00098 0.021 0.0054 1 -0.091 -0.0057 0.00092 0.039
number_prenatal_visits 0.047 -0.13 0.088 -0.036 0.047 0.072 0.047 -0.023 -0.049 -0.091 1 -0.0078 -0.008 0.15
previous_cesarean 0.12 -0.034 0.089 0.0014 -0.053 -0.031 0.12 -0.017 0.0064 0.0057 0.0078 1 5.1e-05 0.02
newborn_gender 0.0041 0.0068 0.0074 0.00084 0.0013 0.026 0.0024 0.0017 0.0022 0.00092 0.008 5.1e-05 1 0.097
newborn_weight 0.061 -0.12 0.17 -0.075 0.14 0.17 0.018 -0.029 -0.061 -0.039 0.15 0.02 0.097 1
mother_body_mass_index
mother_marital_status
mother_delivery_weight
mother_race
mother_height
mother_weight_gain
father_age
father_education
arettes_before_pregnancy
prenatal_care_month
number_prenatal_visits
previous_cesarean
newborn_gender
newborn_weight
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

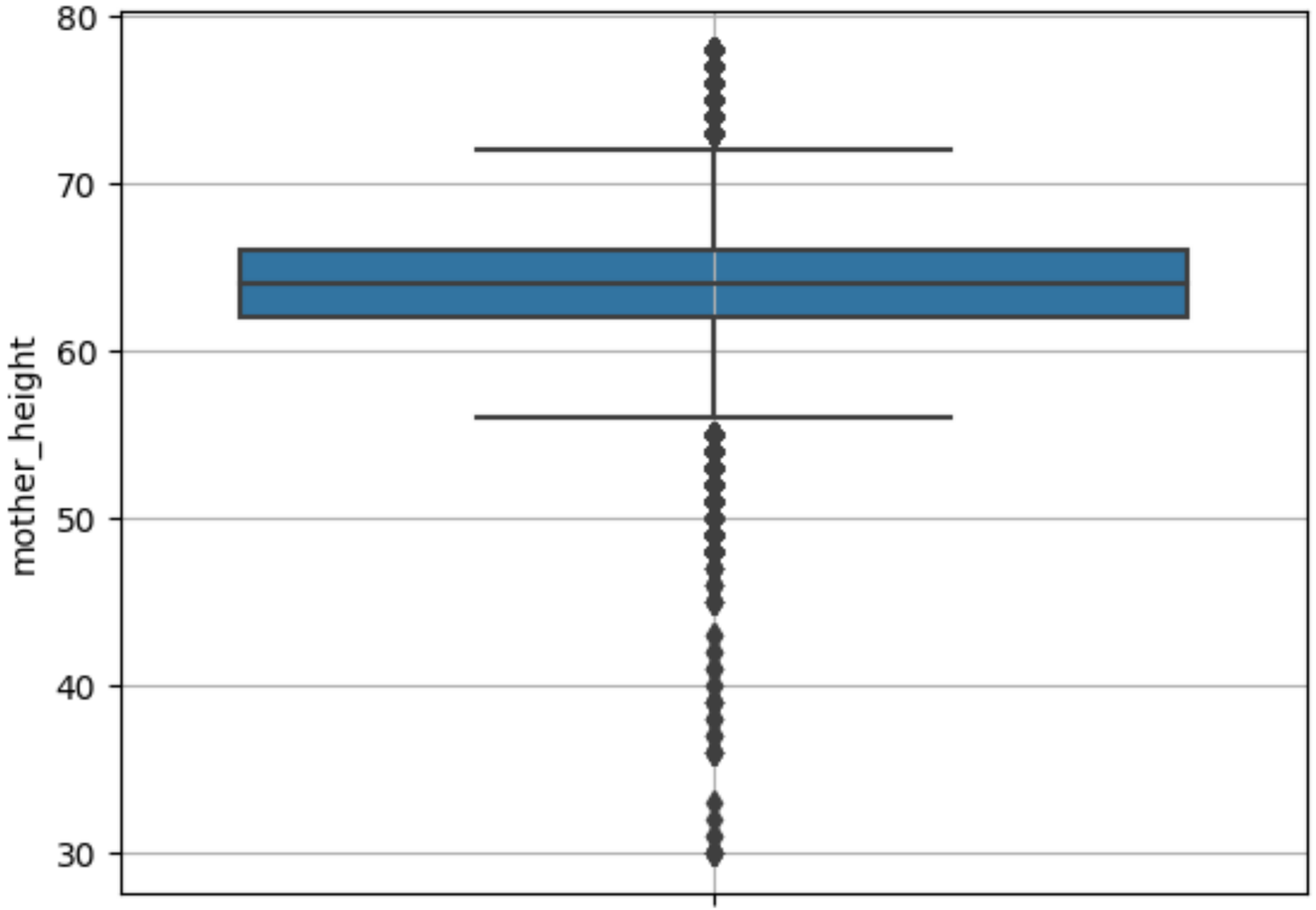
Visualization

Distribution of Discrete Data



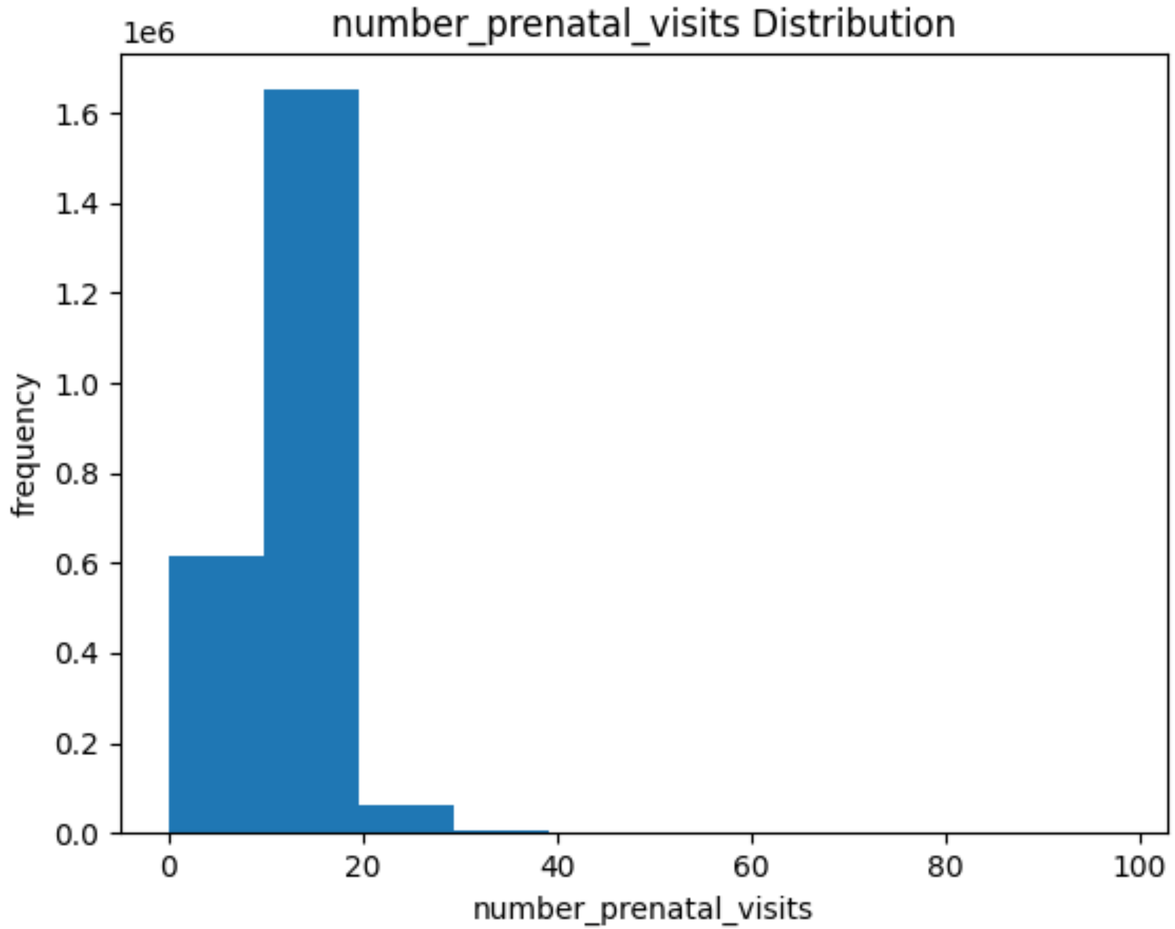
Visualization

Distribution of Float Data



Visualization

Distribution of Int Data



NA Checking

col_name	col_type	NA_ratio
father_age	float64	18.53563
mother_marital_status	float64	17.20142
mother_height	float64	10.19671
mother_body_mass_index	float64	6.11313
mother_weight_gain	float64	3.06378
number_prenatal_visits	float64	2.49784
mother_delivery_weight	float64	1.45773
cigarettes_before_pregnancy	float64	0.47124
mother_race	int64	0.0
father_education	int64	0.0
prenatal_care_month	int64	0.0
previous_cesarean	object	0.0
newborn_gender	object	0.0
newborn_weight	int64	0.0

Feature Engineering

int/float/discrete

- Fill NA/outlier with median/mean to int and float features
- Make a new feature for old with which row has NA
- Truncate outliers
- Standardization and normalization for int and float features
- Make discrete features as dummies
- Power features which have high correlation with newborn weight

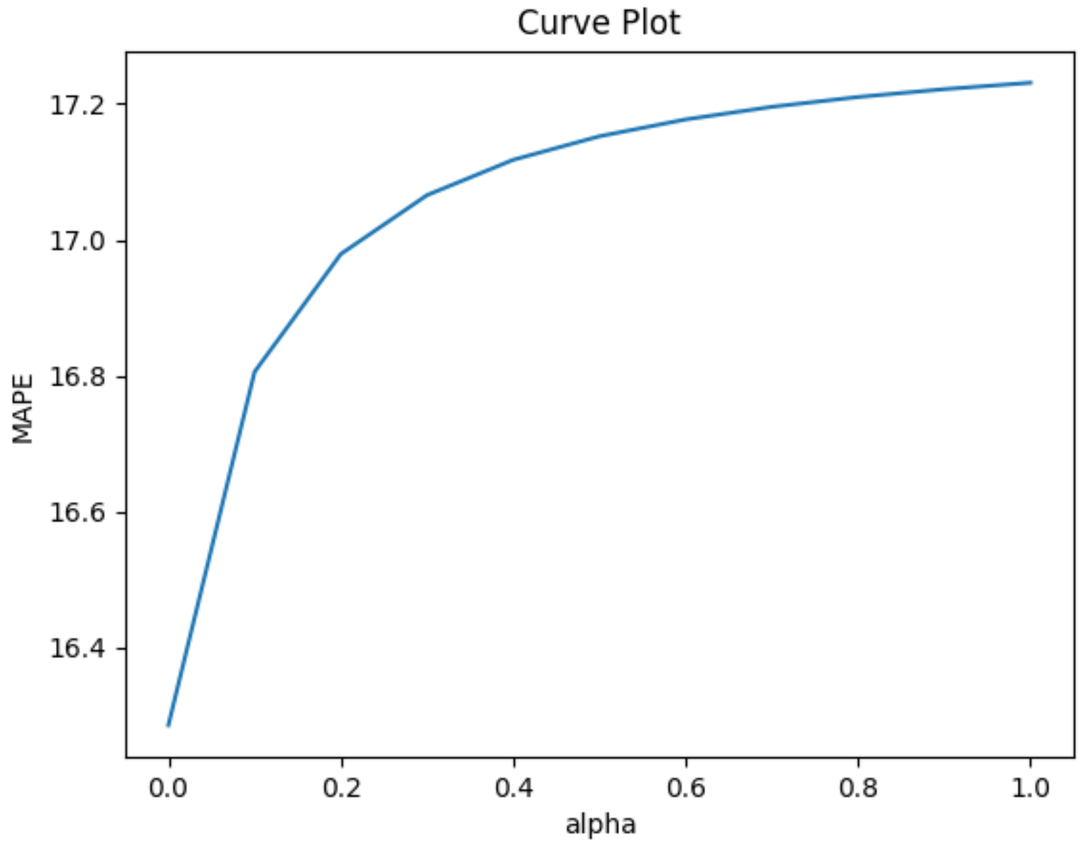
Algorithms

GLM/SVM/GBT

- LM, Ridge, Lasso, ElasticNet, FM
- SVR
- LGBMRegressor

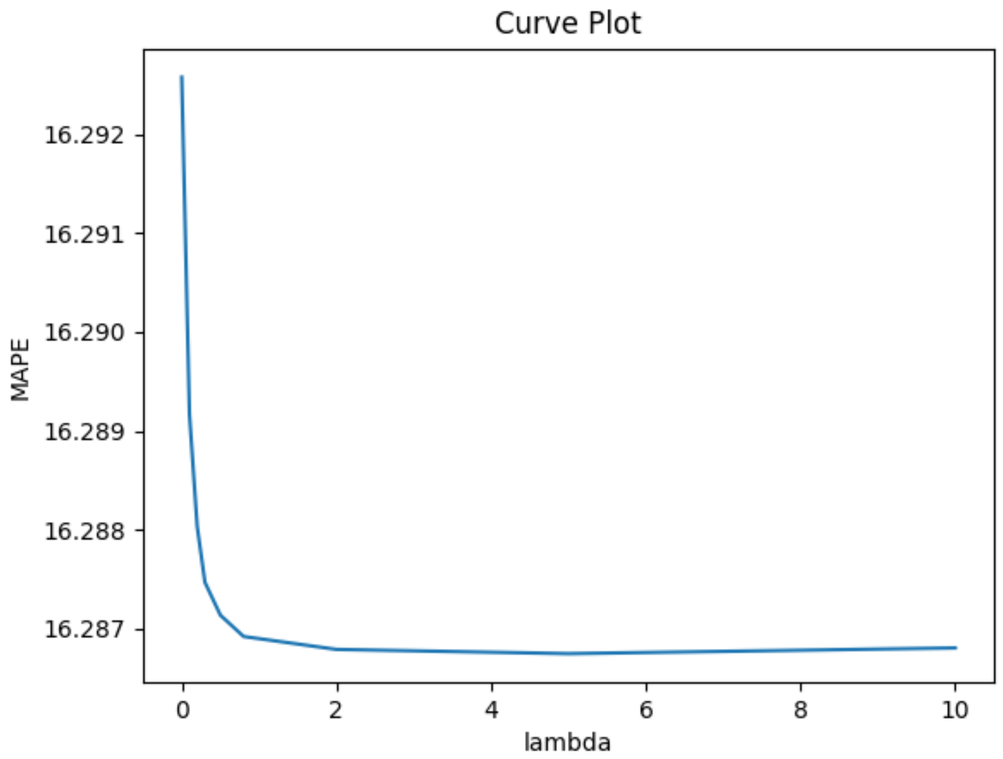
Tuning

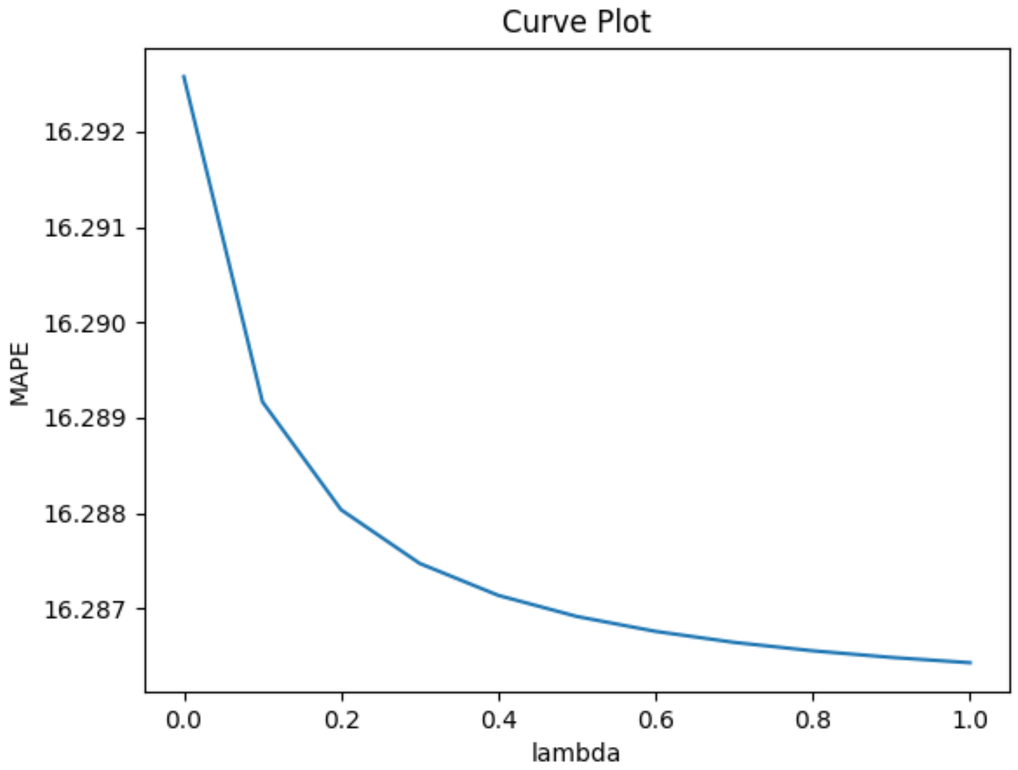
For Lasso



Tuning

For Ridge

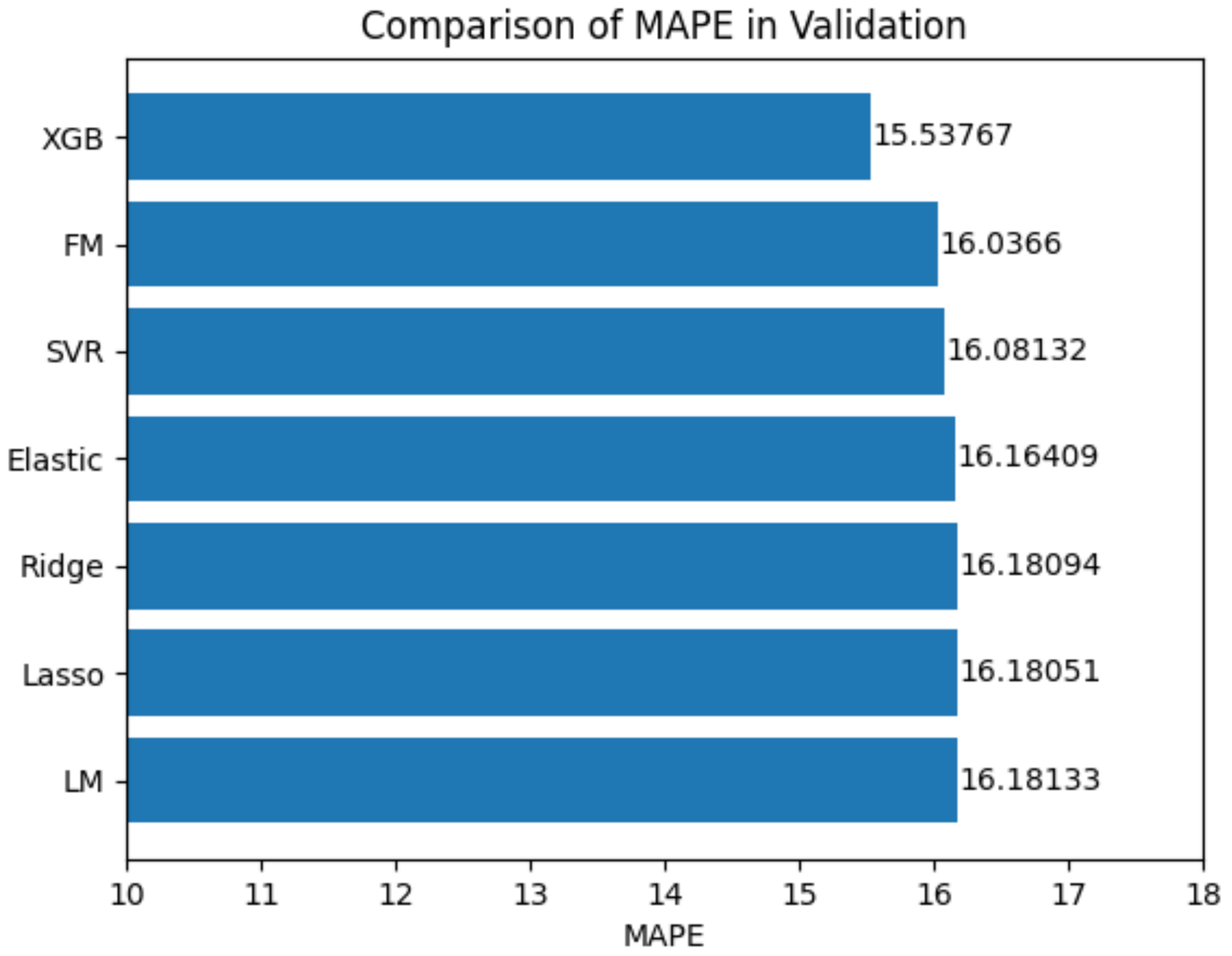


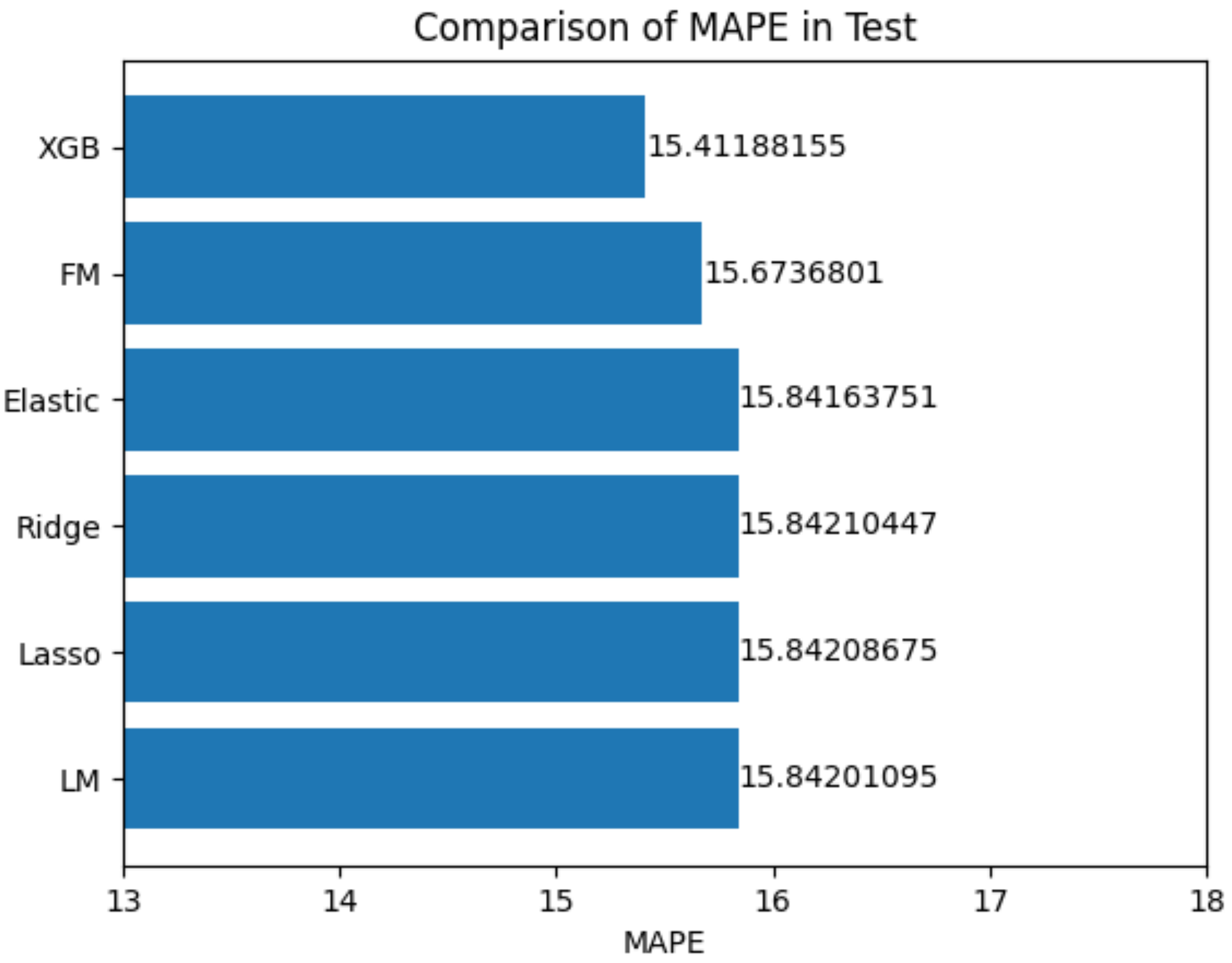


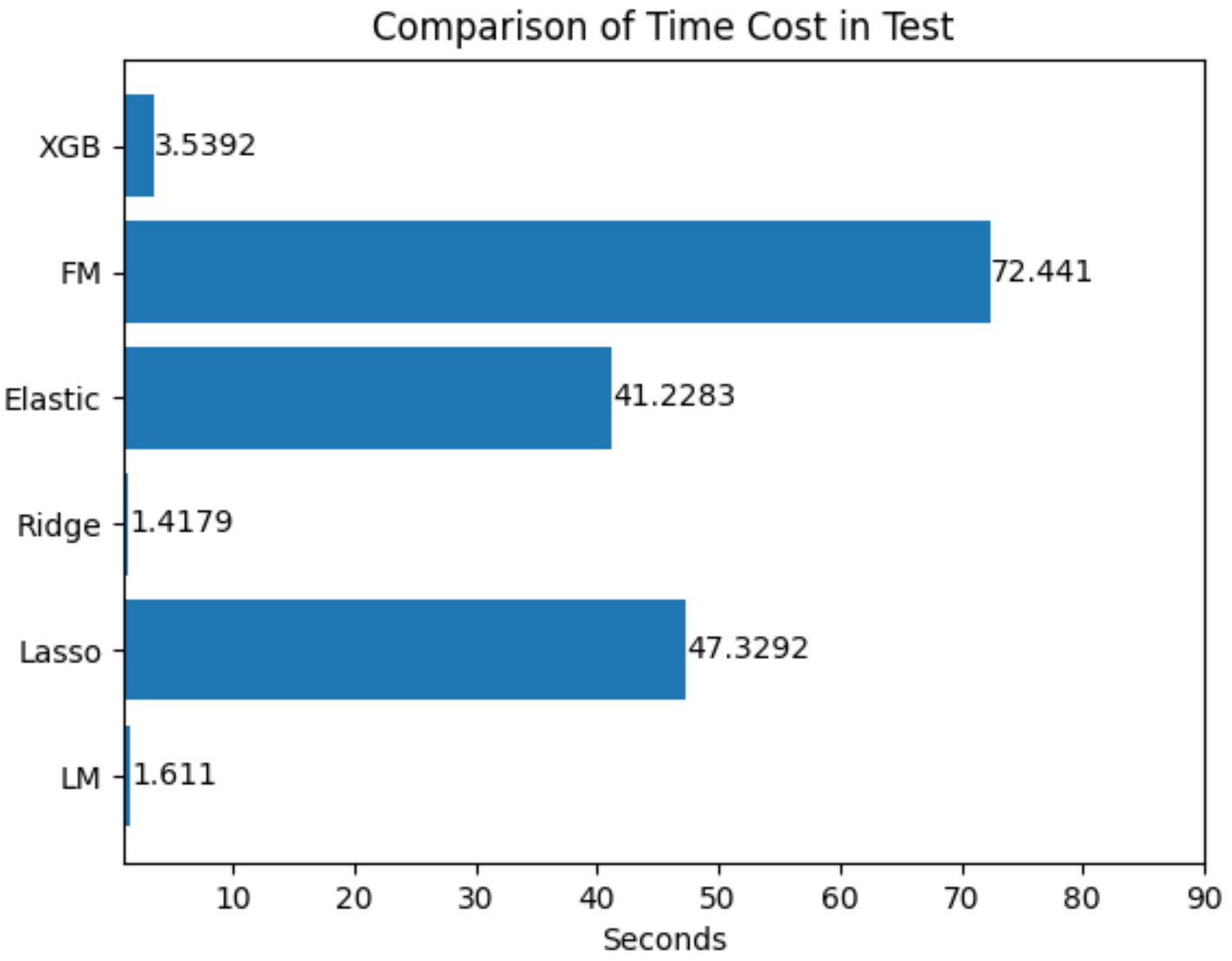
Tuning

For ElasticNet

```
grid_search = GridSearchCV(ElasticNet(),
                           param_grid,
                           cv=5,
                           scoring='mean_absolute_percentage_error')
grid_search.fit(X_train, y_train)
```


Comparison of MAPE in Validation
XGB
FM
SVR
Elastic
Ridge
Lasso
LM
15.53767
16.0366
16.08132
16.16409
16.18094
16.18051
16.18133
10
11
12
13
14
15
16
17
18
MAPE


Comparison of MAPE in Test
XGB
FM
Elastic
Ridge
Lasso
LM
15.41188155
15.6736801
15.84163751
15.84210447
15.84208675
15.84201095
13
14
15
16
17
18
MAPE


Comparison of Time Cost in Test
XGB
FM
Elastic
Ridge
Lasso
LM
3.5392
72.441
41.2283
1.4179
47.3292
1.611
10
20
30
40
50
60
70
80
90
Seconds

Result

LGT

- Train time cost: within 3s for full train data
- MAPE: 15.411% for 10% testset

ML Project for Classification

Yiqing Hu 455858
Raffaele Ricci K-15305

Dimension of dataset = 10127, 21

Na Values

customer sex = 1018
total transaction amount 407
customer age = 624
customer salary range = 681

Unknown Values

customer education = 1519
customer civil status = 749

Feature Selection

#customer_sex

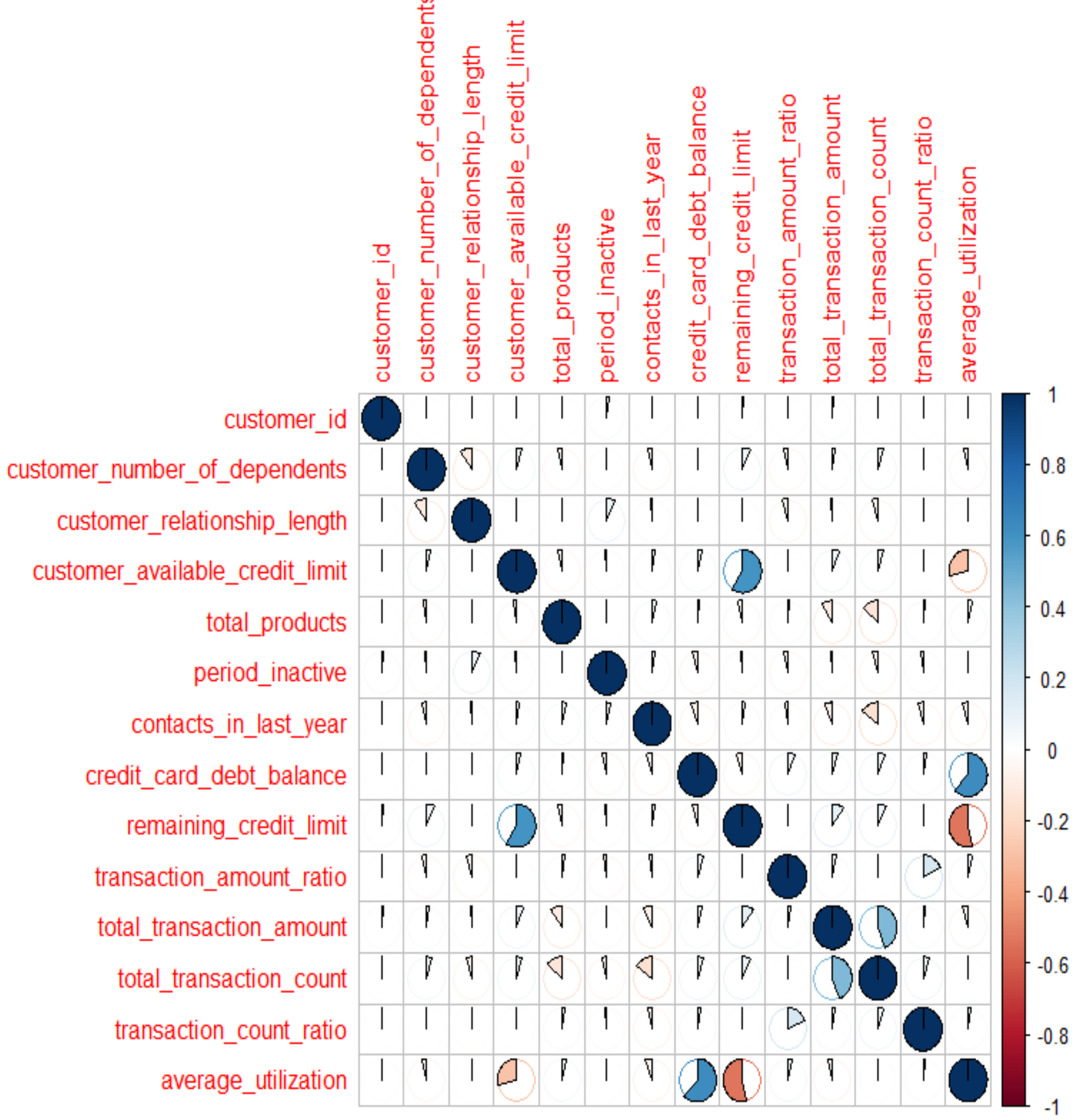
0.003424409 corr. account_status

#customer_age

mean in group closed mean in group open
46.62196 46.25996
Test t: p-value = 0.09492

- we replace the NA value in total transaction amount with the mean
- Replace NA values with "Unknown“
- actually in the variable of salary range we believe that the people with lower or higher salary are unlikely to say own salary

Feature Selection



customer_number_of_dependents	-0.0189905963
customer_relationship_length	-0.0136868512
customer_available_credit_limit	0.0174582135
total_products	0.0795483222
period_inactive	-0.1524488063
contacts_in_last_year	-0.2044905100
credit_card_debt_balance	0.2630528831
remaining_credit_limit	0.0002850775
transaction_amount_ratio	0.1310628478
total_transaction_amount	0.0959313407
total_transaction_count	0.3714027012
transaction_count_ratio	0.1208377823
average_utilization	0.1784103316

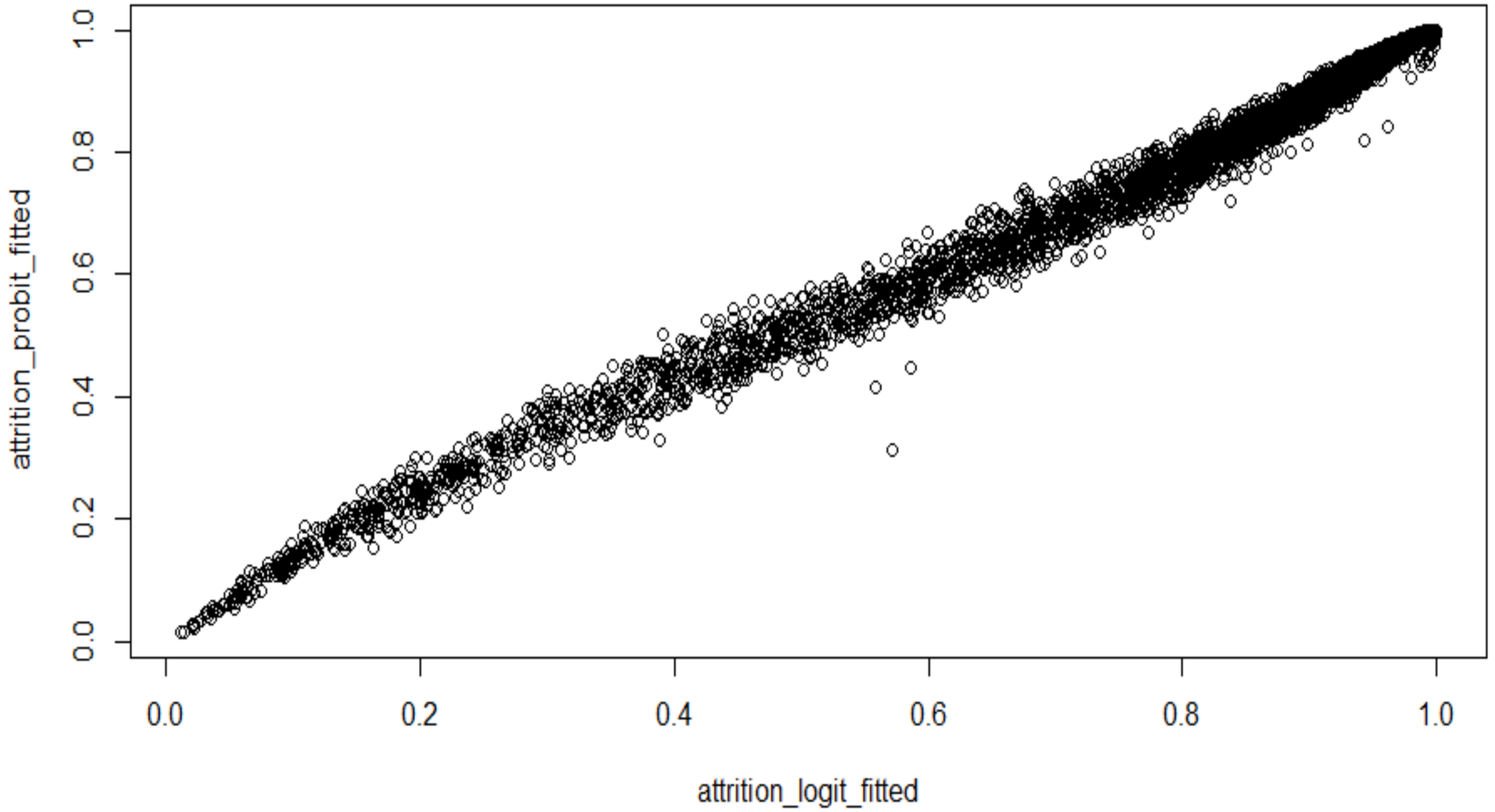
Feature Selection

- Recode ordinal variables with levels, including Unknown as a level

Correlation with account_status
customer_education = 0.035
customer_civil_status = 0.024
customer_salary_range = 0.037
credit_card-classification = 0.015

- we don't remove anything from the earlier apart from the variable sex and age

Logistic Regression



- Division of the dataset in train and test sample
- No big difference in predictions using probit and logit
- Balanced accuracy = 0.735

Feature Selection

In logistic regression we have some insignificant variables

IN LOGIT :

remaining_credit_limit

average_utilization

customer_available_credit_limit

customer_civil_status

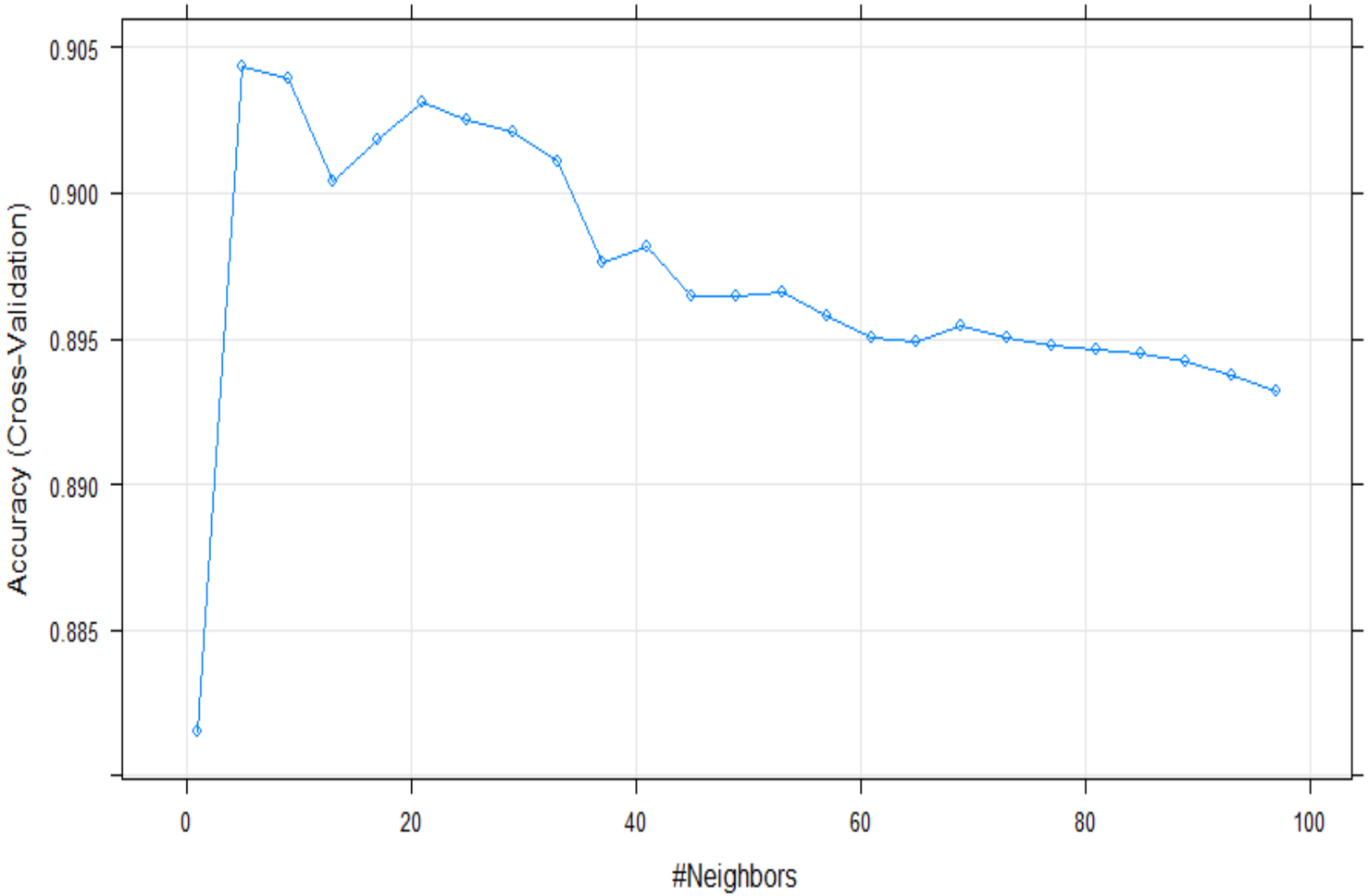
IN PROBIT :

customer_civil_status 4

customer_available_credit_limit 7

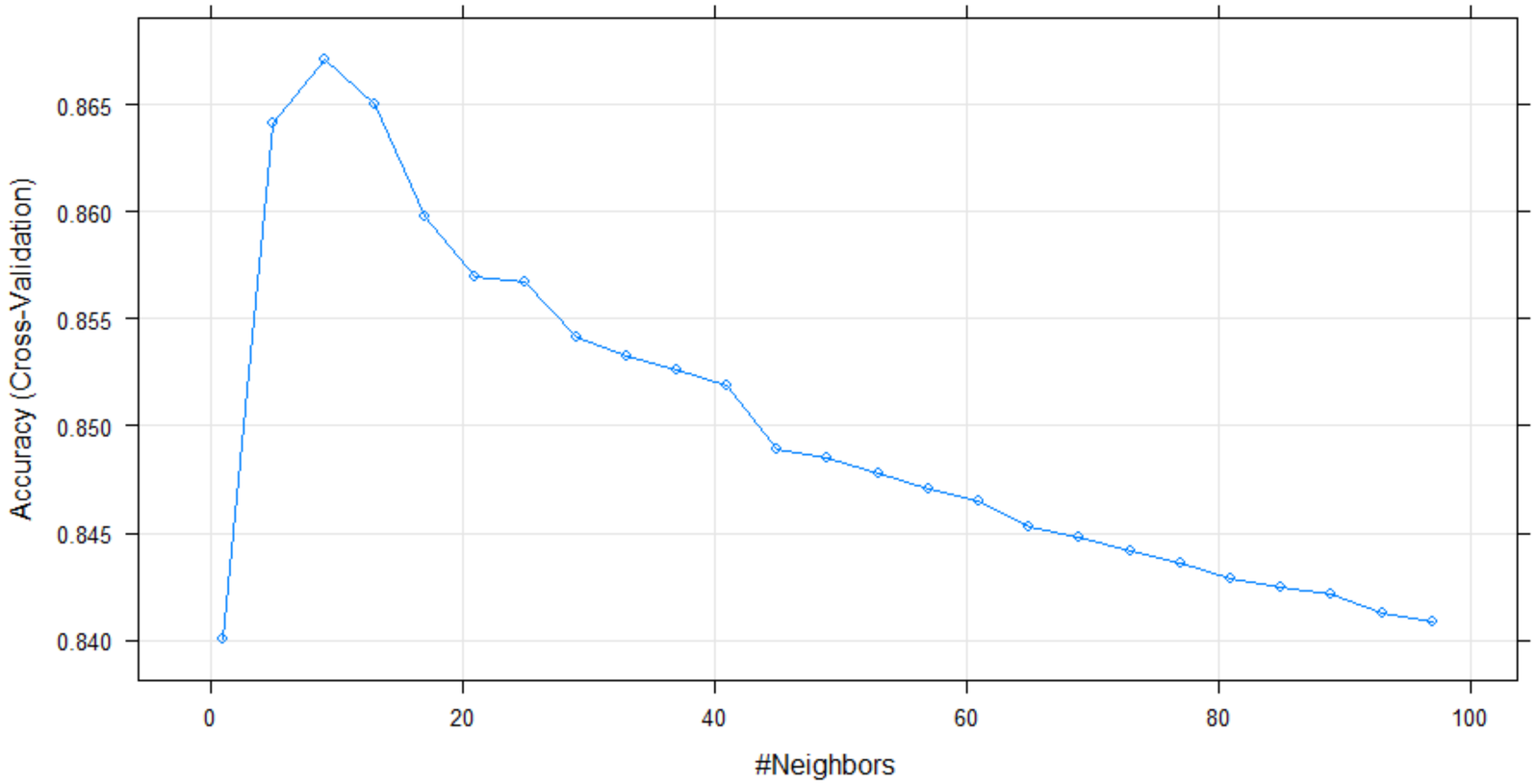
remaining_credit_limit 13

KNN Model



Using cross-validation :
Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
Balanced Accuracy = 0.7995

KNN Model



We compared all the models with scaling the variables but still the best balanced Accuracy remains 0.7995 of the model with K = 5

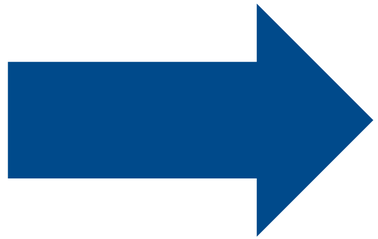
SVM Model

the value of 0.5 is indicated as optimal,
but the accuracy is identical for all C apart from 0.001

# C	Accuracy	Kappa
#0.001	0.8393286	0.0000000
#0.010	0.8918042	0.5004691
0.020	0.8931677	0.5161437
0.050	0.8936379	0.5267727
0.100	0.8938730	0.5299907
0.200	0.8940611	0.5328512
0.500	0.8941081	0.5334885
1.000	0.8940141	0.5330844
2.000	0.8939670	0.5330912
5.000	0.8940611	0.5334981

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was C = 0.5.



Balanced accuracy : 0.73

Thanks guys