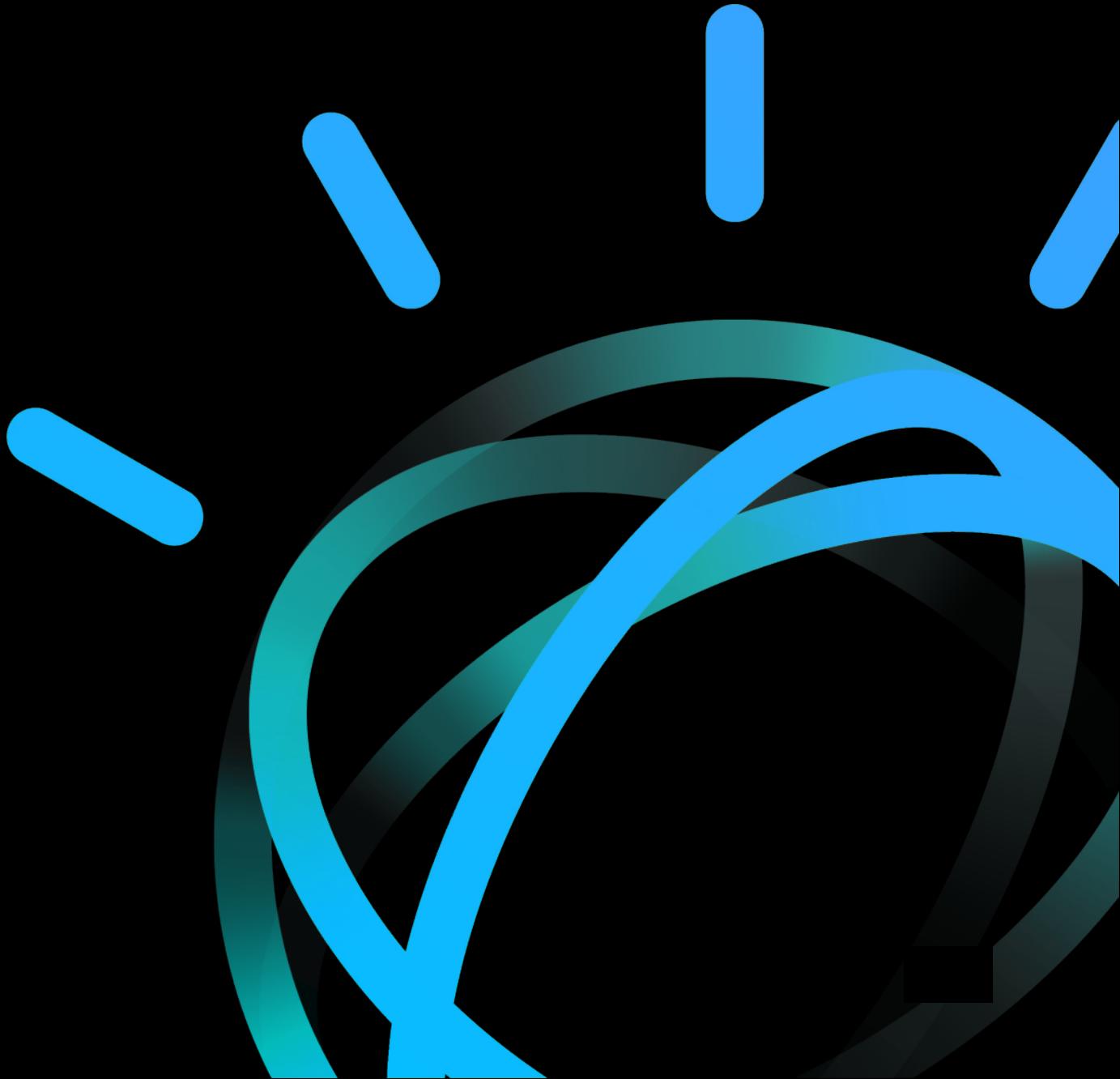


# Capstone Project

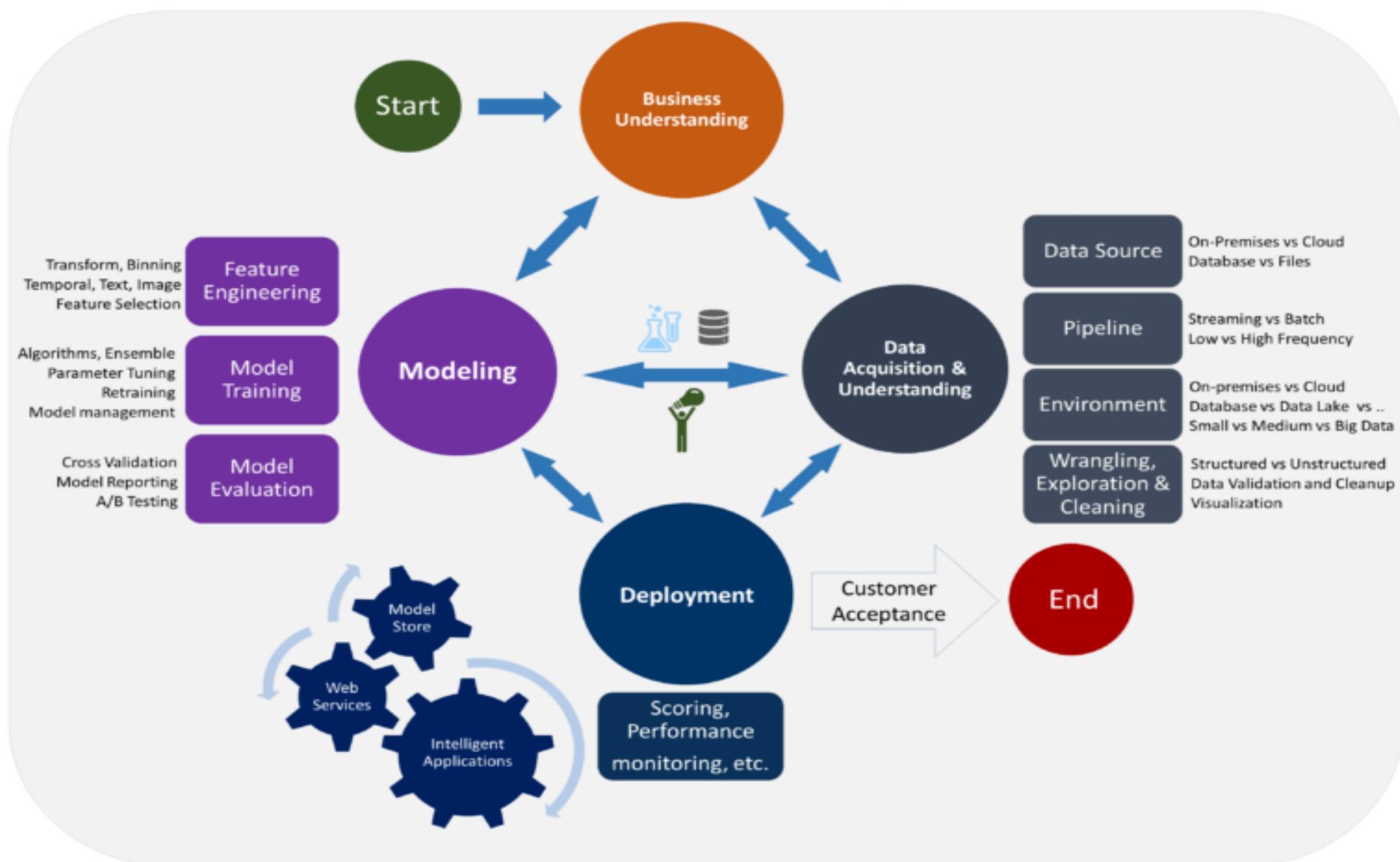
Murat KONCUK



# Agenda

- **Business Understanding**
- **Data Understanding (Exploration & Cleaning)**
- **Modelling**
- **Model Evaluation**
- **Result**

# Data Science Lifecycle



# Business Understanding

	A	B	C	D	E	F	G	H	I	J	K	L
1	Loan Purpose	Checking	Savings	Months Customer	Months Employed	Gender	Marital Status	Age	Housing	Years	Job	Credit Risk
2	9	\$0	\$739	13	12	1	1	23	1	3	1	1
3	3	\$0	\$1.230	25	0	1	2	32	1	1	3	1
4	5	\$0	\$389	19	119	1	1	38	1	4	2	1
5	3	\$638	\$347	13	14	1	1	36	1	2	1	0
6	2	\$963	\$4.754	40	45	1	1	31	2	3	3	0
7	3	\$2.827	\$0	11	13	1	3	25	1	1	3	1
8	5	\$0	\$229	13	16	1	3	26	1	3	1	1
9	1	\$0	\$533	14	2	1	1	27	1	1	1	1
10	9	\$6.509	\$493	37	9	1	1	25	1	2	3	0
11	9	\$966	\$0	25	4	0	2	43	1	1	3	1
12	1	\$0	\$989	49	0	1	1	32	2	2	2	1
13	5	\$0	\$3.305	11	15	1	1	34	2	2	1	1
14	1	\$322	\$578	10	14	1	3	26	1	1	3	1
15	5	\$0	\$821	25	63	1	1	44	1	1	3	1
16	5	\$396	\$228	13	26	1	1	46	1	3	1	0
17	10	\$0	\$129	31	8	1	2	39	1	4	2	1
18	3	\$652	\$732	49	4	0	2	25	1	2	3	0
19	5	\$708	\$683	13	33	1	1	31	1	2	3	1
20	7	\$207	\$0	28	116	1	1	47	1	4	3	1
21	2	\$287	\$12.348	7	2	0	2	23	2	2	3	0
22	3	\$0	\$17.545	34	16	0	2	22	1	4	3	1
23	3	\$101	\$3.871	13	5	0	2	26	2	4	3	0
24	3	\$0	\$0	25	23	1	3	19	1	4	3	1
25	3	\$0	\$485	37	23	0	2	27	1	2	2	1
26	5	\$0	\$10.723	11	15	1	1	39	2	2	1	1
27	1	\$141	\$245	22	33	1	1	26	1	3	3	1
28	10	\$0	\$0	19	58	1	1	50	3	4	3	1
29	10	\$2.484	\$0	49	46	1	1	34	3	1	3	1
30	9	\$237	\$236	37	24	1	1	23	2	4	3	1

# Data Understanding (Statistic)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

## Customer Analytics Classification

Statistics    Summary Graphics    Detailed Graphics    Inferential Statistics    Splitting Dataset    Knn    Decision Tree    SVM    Logistic Regression (GLM)    Naive Bayes

Random Forest    NNET    XGBoost    GBM    LDA    Results

Choose Statistics  
Structure

```
'data.frame': 425 obs. of 12 variables:  
 $ Loan.Purpose : Factor w/ 10 levels "Business","Education",...: 9 3 5 3 2 3 5 1 9 9 ...  
 $ Checking     : num  0 0 0 638 963 ...  
 $ Savings      : num  739 1230 389 347 4754 ...  
 $ Months.Customer: num  13 25 19 13 40 11 13 14 37 25 ...  
 $ Months.Employed: num  12 0 119 14 45 13 16 2 9 4 ...  
 $ Gender       : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 1 ...  
 $ Marital.Status: Factor w/ 3 levels "Divorced","Married",...: 3 1 3 3 3 2 2 3 3 1 ...  
 $ Age          : num  23 32 38 36 31 25 26 27 25 43 ...  
 $ Housing      : Factor w/ 3 levels "Other","Own",...: 2 2 2 2 3 2 2 2 2 2 ...  
 $ Years         : num  3 1 4 2 3 1 3 1 2 1 ...  
 $ Job           : Factor w/ 3 levels "Management","Skilled",...: 3 2 1 3 2 2 3 3 2 2 ...  
 $ Credit.Risk   : Factor w/ 2 levels "High","Low": 1 1 1 2 2 1 1 1 2 1 ...'
```

Loan.Purpose	Checking	Savings	Months.Customer	Months.Employed	Gender	Marital.Status	Age	Housing	Years	Job	Credit.Risk
Small Appliance	0	739	13	12	M	Single	23	Own	3	Unskilled	High
Furniture	0	1230	25	0	M	Divorced	32	Own	1	Skilled	High
New Car	0	389	19	119	M	Single	38	Own	4	Management	High
Furniture	638	347	13	14	M	Single	36	Own	2	Unskilled	Low
Education	963	4754	40	45	M	Single	31	Rent	3	Skilled	Low
Furniture	2827	0	11	13	M	Married	25	Own	1	Skilled	High

Ubuntu icon    Home icon    Terminal icon    File icon    Google icon    R icon    Battery icon    Bluetooth icon    WiFi icon    Volume icon    Date/Time: 08:19

# Data Understanding (Statistic)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

## Customer Analytics Classification

Statistics    Summary Graphics    Detailed Graphics    Inferential Statistics    Splitting Dataset    Knn    Decision Tree    SVM    Logistic Regression (GLM)    Naive Bayes

Random Forest    NNET    XGBoost    GBM    LDA    Results

**Choose Statistics**

Summary

Loan.Purpose	Checking	Savings	Months.Customer
Small Appliance:105	Min. : 0	Min. : 0	Min. : 5.0
New Car :104	1st Qu.: 0	1st Qu.: 228	1st Qu.:13.0
Furniture : 85	Median : 0	Median : 596	Median :19.0
Business : 44	Mean : 1048	Mean : 1813	Mean :22.9
Used Car : 40	3rd Qu.: 560	3rd Qu.: 921	3rd Qu.:28.0
Education : 23	Max. :19812	Max. :19811	Max. :73.0
(Other) : 24			

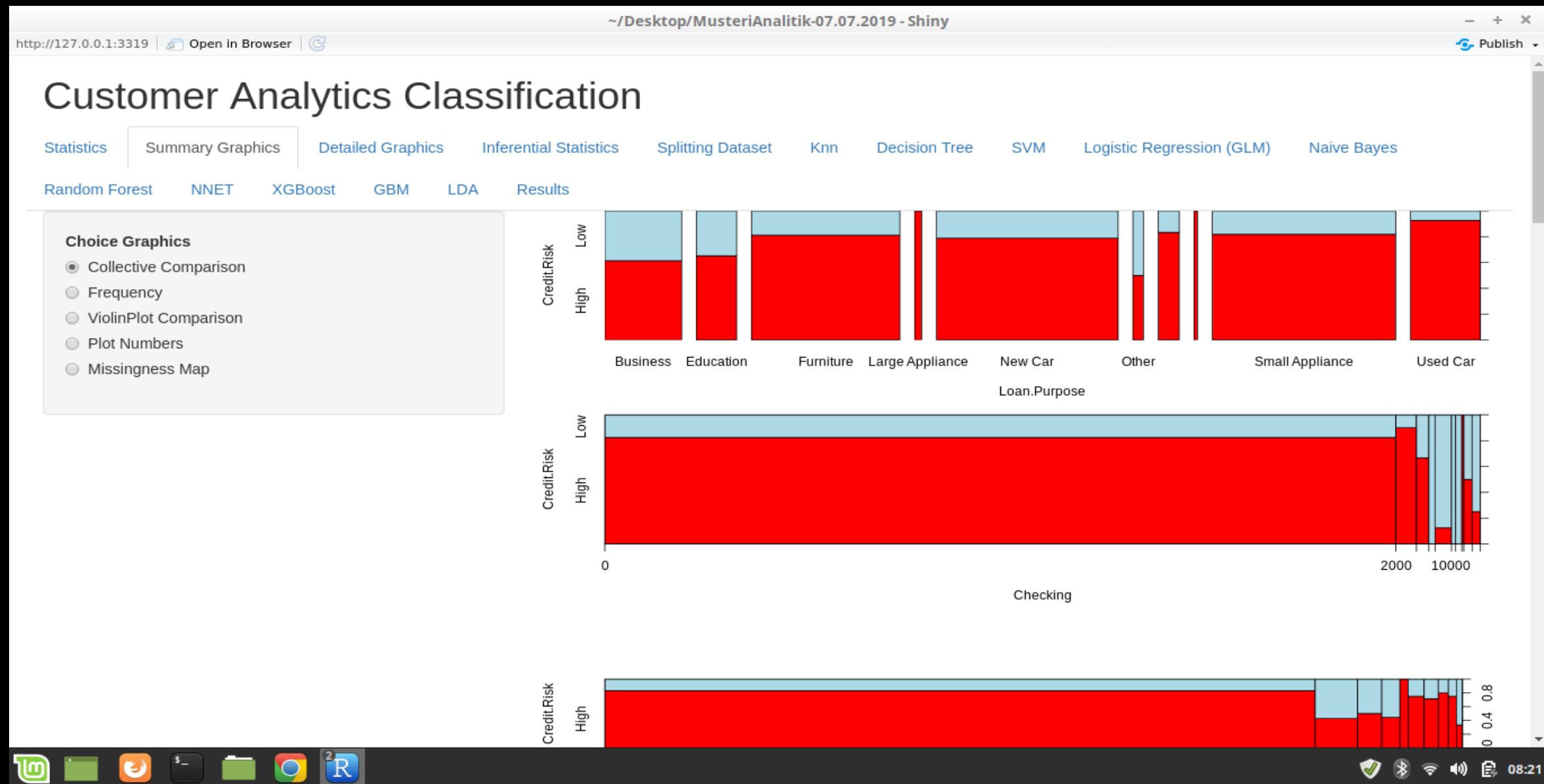
Months.Employed	Gender	Marital.Status	Age	Housing
Min. : 0.0	F:135	Divorced:156	Min. :18.0	Other: 52
1st Qu.: 6.0	M:290	Married : 36	1st Qu.:26.0	Own :292
Median : 20.0		Single :233	Median :32.0	Rent : 81
Mean : 31.9			Mean :34.4	
3rd Qu.: 47.0			3rd Qu.:41.0	
Max. :119.0			Max. :73.0	

Years	Job	Credit.Risk
Min. :1.00	Management: 54	High:335
1st Qu.:2.00	Skilled :271	Low : 90
Median :3.00	Unskilled :100	
Mean : 2.84		
3rd Qu.:4.00		
Max. :4.00		

Loan.Purpose	Checking	Savings	Months.Customer	Months.Employed	Gender	Marital.Status	Age	Housing	Years	Job	Credit.Risk
Small	0	739	13	12	M	Single	23	Own	3	Unskilled	High

Icons: R, GitHub, File, Folder, Home, Stop, Refresh, Help, 08:20

# Data Understanding (Visualize)



# Data Understanding (Visualize)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser |

## Customer Analytics Classification

Statistics    Summary Graphics    Detailed Graphics    Inferential Statistics    Splitting Dataset    Knn    Decision Tree    SVM    Logistic Regression (GLM)    Naive Bayes

Random Forest    NNET    XGBoost    GBM    LDA    Results

**Choice Graphics**

- Collective Comparison
- Frequency
- ViolinPlot Comparison
- Plot Numbers
- Missingness Map

Results

Bar chart showing the frequency or percentage of various purposes:

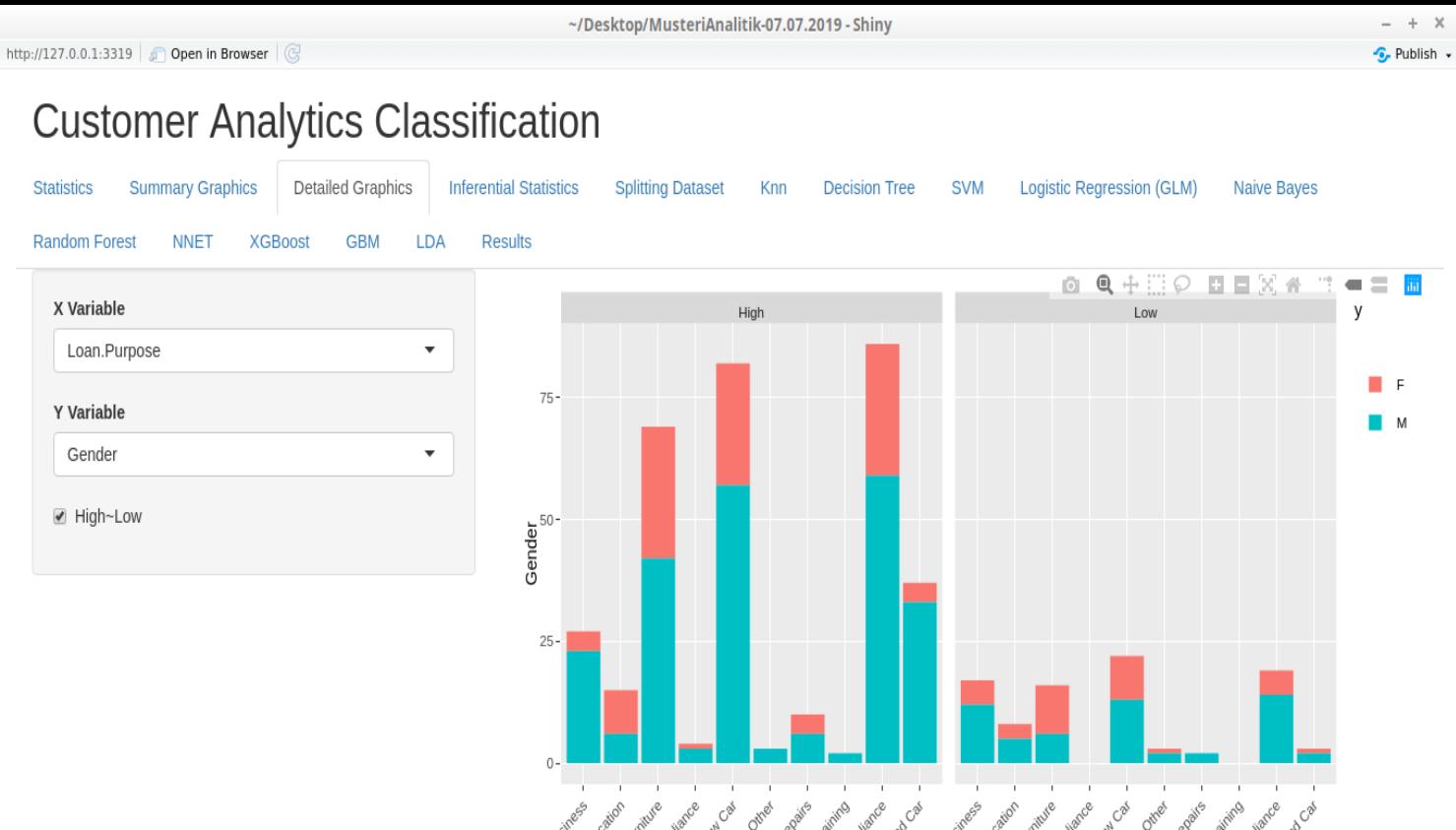
Purpose	Frequency / (Percentage %)
Small Appliance	105 (24.71%)
New Car	104 (24.47%)
Furniture	85 (20%)
Business	44 (10.35%)
Used Car	40 (9.41%)
Education	23 (5.41%)
Repairs	12 (2.82%)
Other	6 (1.41%)
Large Appliance	4 (0.94%)
Retraining	2 (0.47%)

Bar chart showing the frequency or percentage of the letter M:

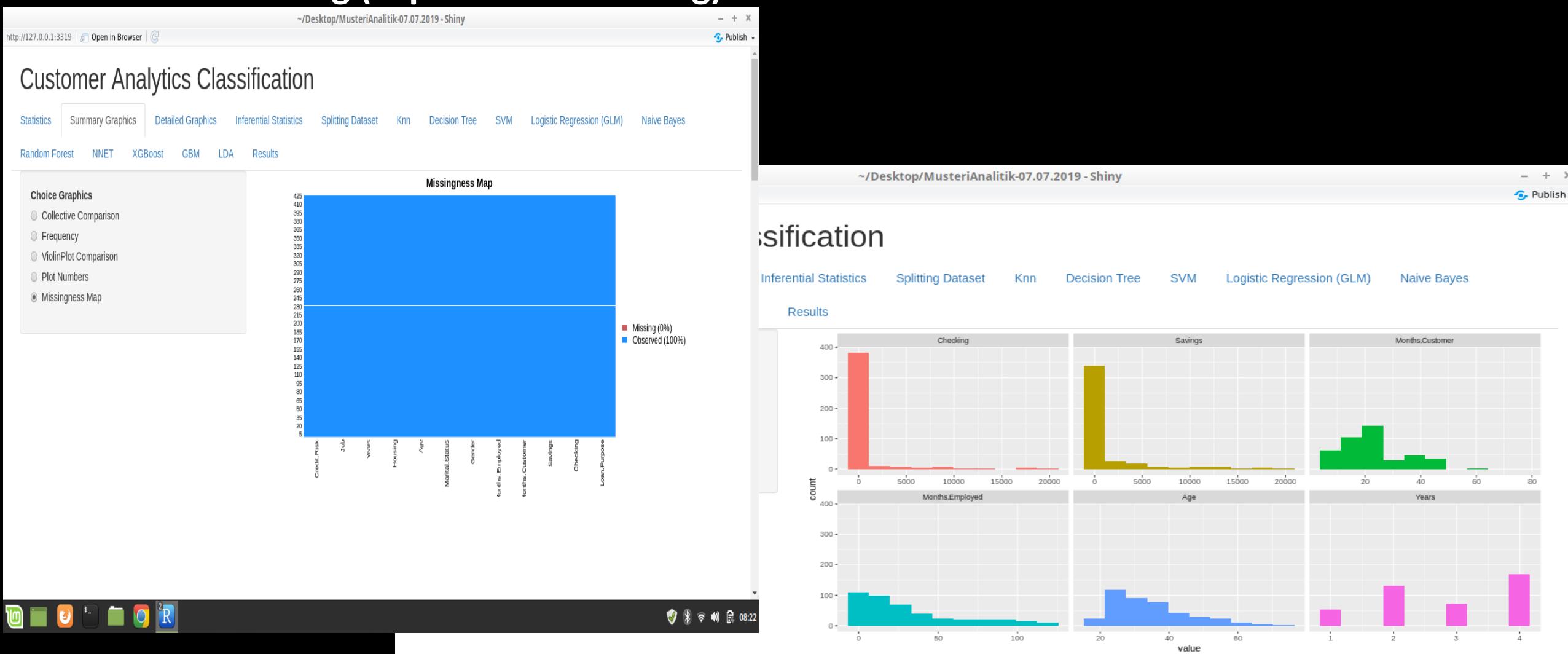
M	Frequency / (Percentage %)
M	290 (68.24%)

System tray icons: 08:21

# Data Understanding (Exploration-Cleaning)



# Data Understanding (Exploration-Cleaning)



08:22

08:22

# Data Understanding (Exploration-Cleaning-Python)

```
In [10]: cmap = 'coolwarm'  
corr=df.corr()  
corr.style.background_gradient(cmap, axis=1)
```

Out[10]:

	Checking	Savings	Months Customer	Months Employed	Gender	Marital Status	Age	Housing	Years	Job	Credit Risk
Checking	1	0.0200792	-0.0362263	-0.00919098	0.0207947	-0.0175604	-0.0023693	0.0396903	-0.0355347	0.0067802	-0.351463
Savings	0.0200792	1	-0.0504407	0.048978	-0.04985	0.0251077	-0.0282859	-0.0422778	-0.0193616	0.0186301	-0.141489
Months Customer	-0.0362263	-0.0504407	1	0.0544558	0.125811	-0.133166	-0.00378591	0.220757	0.0831089	0.16697	-0.0838113
Months Employed	-0.00919098	0.048978	0.0544558	1	0.246575	-0.259574	0.306799	0.0486035	0.28823	0.0113563	0.081447
Gender	0.0207947	-0.04985	0.125811	0.246575	1	-0.489119	0.156988	-0.0449949	-0.0214099	-0.0414339	0.066945
Marital Status	-0.0175604	0.0251077	-0.133166	-0.259574	-0.489119	1	-0.221221	-0.0948081	-0.149197	0.0243386	-0.131088
Age	-0.0023693	-0.0282859	-0.00378591	0.306799	0.156988	-0.221221	1	0.094231	0.240027	-0.134193	0.0479164
Housing	0.0396903	-0.0422778	0.220757	0.0486035	-0.0449949	-0.0948081	0.094231	1	0.335998	0.049325	-0.0314407
Years	-0.0355347	-0.0193616	0.0831089	0.28823	-0.0214099	-0.149197	0.240027	0.335998	1	-0.0348175	0.0933453
Job	0.0067802	0.0186301	0.16697	0.0113563	-0.0414339	0.0243386	-0.134193	0.049325	-0.0348175	1	0.0625659

```
Out[9]: array([[ 1.14004792, -0.3333931 , -0.29878856, ..., -0.62151966,  
   0.14734777, -1.78085412],  
  [-0.84704859, -0.3333931 , -0.16213587, ..., -0.62151966,  
   -1.6944994 ,  0.6685407 ],  
  [-0.18468309, -0.3333931 , -0.39619882, ..., -0.62151966,  
   1.06827136, -0.55615671],  
  ...,  
  [-0.18468309, -0.3333931 , -0.50446338, ..., -0.62151966,  
   -0.77357581,  0.6685407 ],  
  [-0.18468309, -0.3333931 , -0.30630306, ..., -0.62151966,  
   -0.77357581,  0.6685407 ],  
  [-0.18468309, -0.3333931 , -0.25064005, ..., -0.62151966,  
   0.14734777, -0.55615671]])
```

# Modelling (Data Spilit)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | C | Publish ▾

## Customer Analytics Classification

Statistics    Summary Graphics    Detailed Graphics    Inferential Statistics    **Splitting Dataset**    Knn    Decision Tree    SVM    Logistic Regression (GLM)    Naive Bayes

Random Forest    NNET    XGBoost    GBM    LDA    Results

Percentage of Test Set:



The Size of DataSet : 425  
Test Set : 106  
Train Set : 319

U G T \$ F R 08:25

# Modelling (Inferential)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

## Customer Analytics Classification

Statistics    Summary Graphics    Detailed Graphics    Inferential Statistics    Splitting Dataset    Knn    Decision Tree    SVM    Logistic Regression (GLM)    Naive Bayes

Random Forest    NNET    XGBoost    GBM    LDA    Results

### Wilcoxon Rank Sum and Signed Rank Tests

Between High and Low Credit Risk Classes within Variables

(None of the variables have a normal distribution according to the Shapiro–Wilk test)

Comparison of P Value

Variable	P_Value	statically
Loan.Purpose	0.0021350000	Meaningful P<0.05
Checking	0.0000000000	Meaningful P<0.05
Savings	0.0000000000	Meaningful P<0.05
Months.Customer	0.4226670000	
Months.Employed	0.2749990000	
Gender	0.1682390000	
Marital.Status	0.0057930000	Meaningful P<0.05
Age	0.1104190000	
Housing	0.3235830000	
Years	0.0475830000	Meaningful P<0.05
Job	0.1697820000	

Shapiro-Wilk Test Results

Variables	P_Value	statically
Loan.Purpose	0.0000000000	Not Normal P<0.05
Checking	0.0000000000	Not Normal P<0.05
Savings	0.0000000000	Not Normal P<0.05
Months.Customer	0.0000000000	Not Normal P<0.05
Months.Employed	0.0000000000	Not Normal P<0.05
Gender	0.0000000000	Not Normal P<0.05
Marital.Status	0.0000000000	Not Normal P<0.05
Age	0.0000000000	Not Normal P<0.05
Housing	0.0000000000	Not Normal P<0.05
Years	0.0000000000	Not Normal P<0.05
Job	0.0000000000	Not Normal P<0.05

U    G    E    \$    F    Google Chrome    R

✓    Bluetooth    WiFi    Sound    08:24

# Modelling (Decision Tree)

Customer Analytics Classification

Statistics   Summary Graphics   Detailed Graphics   Inferential Statistics   Splitting Dataset   Knn   Decision Tree   SVM   Logistic Regression (GLM)   Naive Bayes

Random Forest   NNET   XGBoost   GBM   LDA   Results

Conditional Inference Trees Algorithm

Accuracy Rate : 0.92

Prediction	Reference	Freq
0	0	19
1	0	3
0	1	5
1	1	79

Gender (0-1)

Loan Purpose (1-2)

Checking

Savings

Months Customer

Months Employee

Decision Tree Diagram:

```
graph TD; Node1((1)) -- "Checking p < 0.001" --> Node2((2)); Node1 -- "Savings p = 0.001" --> Node5((5)); Node2 -- "<= 256" --> Node3((Node 3 n = 189)); Node2 -- "> 256" --> Node5; Node3 -- "0" --> Node3a[0]; Node3 -- "1" --> Node3b[1]; Node4((Node 4 n = 22)) --- Node3b; Node5 -- "<= 124" --> Node6((Node 6 n = 28)); Node5 -- "> 124" --> Node7((Node 7 n = 80)); Node6 --- Node6a[0]; Node6 --- Node6b[1]; Node7 --- Node7a[0]; Node7 --- Node7b[1];
```

Confusion Matrix and Statistics

Confusion Matrix and Statistics	
Reference	Prediction
0	1
0	19
1	5
1	3
1	79

Accuracy : 0.925  
95% CI : (0.857, 0.967)

# Modelling (Decision Tree-Python)

```
In [55]: credittree.fit(X_train,y_train)  
Out[55]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=4,  
max_features=None, max_leaf_nodes=None,  
min_impurity_decrease=0.0, min_impurity_split=None,  
min_samples_leaf=1, min_samples_split=2,  
min_weight_fraction_leaf=0.0, presort=False, random_state=None,  
splitter='best')
```

```
In [56]: predTree = credittree.predict(X_test)
```

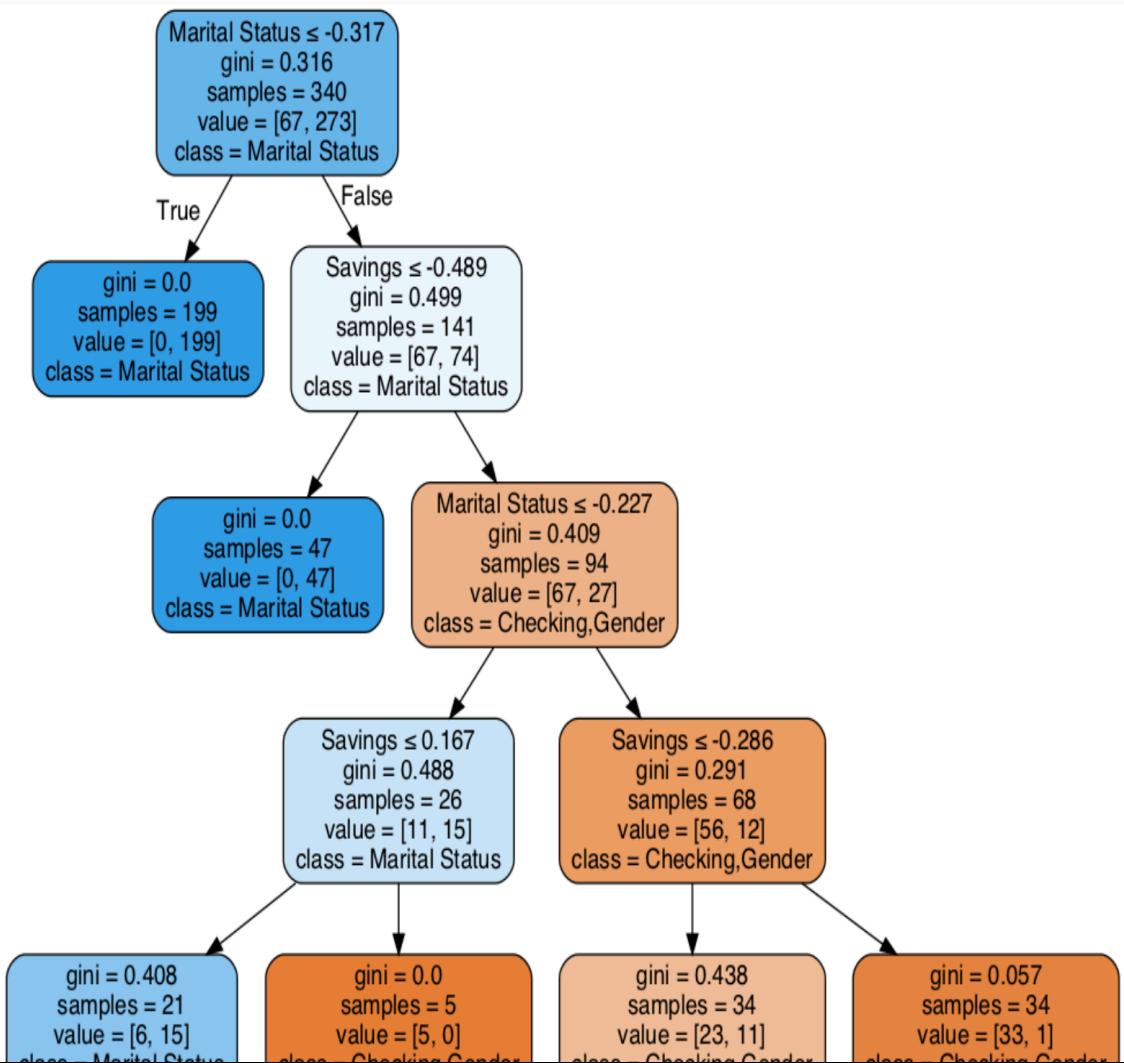
```
In [57]: print (predTree [0:12])  
print (y_test [0:12])
```

```
[0. 0. 0. 1. 1. 1. 1. 1. 1. 1.]  
[1. 1. 0. 1. 1. 1. 1. 1. 1. 1.]
```

```
In [58]: from sklearn import metrics  
import matplotlib.pyplot as plt  
print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_test, predTree))  
  
DecisionTrees's Accuracy: 0.9058823529411765
```

```
In [59]: from sklearn.tree import export_graphviz  
from sklearn.externals.six import StringIO  
from IPython.display import Image  
import pydotplus  
  
col = ["Checking", "Gender", "Marital Status", "Savings", "Months Customer", "Months Employed", "Marital Status", "Age", "Hous  
dot_data = StringIO()  
export_graphviz(credittree, out_file=dot_data,  
filled=True, rounded=True,  
special_characters=True, feature_names = col, class_names = col)  
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())  
graph.write_png('decision.png')  
Image(graph.create_png())
```

Out[59]:



# Modelling (GBM-GLM)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser |

Gradient Boosting Machines Model  
shrinkage (learning rate)  
0.06

cv.folds  
2 6 20

n.trees (Number of trees)  
502 10,000

Distributions  
adaboost

Accuracy Rate :  
Using 99 trees... 0.97

Prediction Reference Freq

0	0	21
1	0	1
0	1	2
1	1	82

Confusion Matrix

Confusion Matrix

ROC

Gender (0-1)

Loan Purpose (1-10)

Marital Status (1-3)

Checking

Age

Savings

Months Customer

Years(1-4)

Months Employed

Job(1-3)

Customer Analytics Classification

Statistics Summary Graphics Detailed Graphics Inferential Statistics Splitting Dataset Knn Decision Tree SVM Logistic Regression (GLM) Naive Bayes

Random Forest NNET XGBoost GBM LDA Results

Generalized Linear Model

Accuracy Rate : 0.79

Prediction Reference Freq

0	0	6
1	0	16
0	1	6
1	1	78

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser |

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser |

Logout

File Edit View Insert Cell Help

## Modelling (KNN-LDA )

Customer Analytics Classification

Statistics    Summary Graphics    Detailed Graphics    Inferential Statistics    Splitting Dataset    Knn    Decision Tree    SVM    Logistic Regression (GLM)    Naive Bayes

Random Forest    NNET    XGBoost    GBM    LDA    Results

K Nearest Neighbors Algorithm

Number of Neighbors

Accuracy Rate : 0.89

Prediction	Reference	Freq
0	0	16
1	0	6
0	1	6
1	1	78

Confusion Matrix and Statistics

Confusion Matrix and Statistics

Gender (0-1)

Loan Purpose

Checking

Savings

Months Customer

Months Employed

Linear Discriminant Analysis

Accuracy Rate :

0.82

Prediction	Reference	Freq
0	0	6
1	0	16
0	1	3
1	1	81

Confusion Matrix and Statistics

Confusion Matrix and Statistics

Gender (0-1)

Loan Purpose (1-10)

Marital Status (1-3)

Checking

Age

Savings

Housing (1-3)

Months Customer

Years(1-4)

Months Employed

Job(1-3)

# Modelling (NaiveBayes - NNET)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

## Customer Analytics Classification

Statistics   Summary Graphics   Detailed Graphics   Inferential Statistics   Splitting Dataset   Knn   Decision Tree   SVM   Logistic Regression (GLM)   **Naive Bayes**

Random Forest   NNET   XGBoost   GBM   LDA   Results

Naive Bayes

Accuracy Rate : 0.81

Prediction	Reference	Freq
0	0	7
1	0	15
0	1	5
1	1	79

Gender (0-1)

Loan Purpose (1-1)

Checking

Savings

Months Customer

Months Employed

Job(1-3)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

## Accuracy Rate :

0.83

Prediction	Reference	Freq
0	0	8
1	0	14
0	1	4
1	1	80

Months Employed

Job(1-3)

Loan.Purpose

Checking

Savings

Months.Customer

Months.Employed

Gender

Marital.Status

Age

Housing

Years

Job

B1

B2

O1 Credit.Risk

H1

H2

H3

H4

ROC



Confusion Matrix and Statistics

Confusion Matrix and Statistics

Reference	0	1
Prediction	0	8
0	8	4



# Modelling (Random Forest-SVM)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

## Customer Analytics Classification

Statistics Summary Graphics Detailed Graphics Inferential Statistics Splitting Dataset Knn Decision Tree SVM Logistic Regression (GLM) Naive Bayes

Random Forest NNET XGBoost GBM LDA Results

Random Forest  
Number of Trees:  
10 1,000 2,000

Accuracy Rate :  
0.95

Prediction	Reference	Freq
0	0	18
1	0	4
0	1	1
1	1	83

Gender (0-1)  
Loan Purpose (1-10)  
Checking  
Savings  
Months Customer  
Months Employed

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

## Customer Analytics Classification

Statistics Summary Graphics Detailed Graphics Inferential Statistics Splitting Dataset Knn Decision Tree SVM Logistic Regression (GLM) Naive Bayes

Random Forest NNET XGBoost GBM LDA Results

Support Vector Machine Algorithm  
Chose Kernel  
radial

Accuracy Rate :  
0.82

Prediction	Reference	Freq
0	0	6
1	0	16
0	1	3
1	1	81

Gender (0-1)  
Loan Purpose (1-10)  
Marital Status (1-3)  
Checking  
Age  
Savings  
Housing (1-3)  
Months Customer  
Years(1-4)  
Months Employed  
Job(1-3)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

Ubuntu icon bar: Home, Applications, Dash, RStudio, File Manager, Google Chrome, RStudio Server.

System tray icons: Network, Bluetooth, WiFi, Volume, Battery, Date/Time (08:26).

# Modelling (XGBOOST )

~ /Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser |

Extreme Gradient Boosting

Max.Depth(Tree)

eta (learning rate)

nthread (Number of parallel threads)

nrounds (Number of Iterations)

Predict Status

Gender (0-1)

Loan Purpose (1-10)

Marital Status (1-3)

Checking

Age

Savings

Housing (1-3)

Months Customer

Years(1-4)

Months Employed

Job(1-3)

Accuracy Rate :

0.96

Confusion Matrix and Statistics

Confusion Matrix and Statistics

Prediction	Reference	Freq
0	0	20
1	0	2
0	1	2

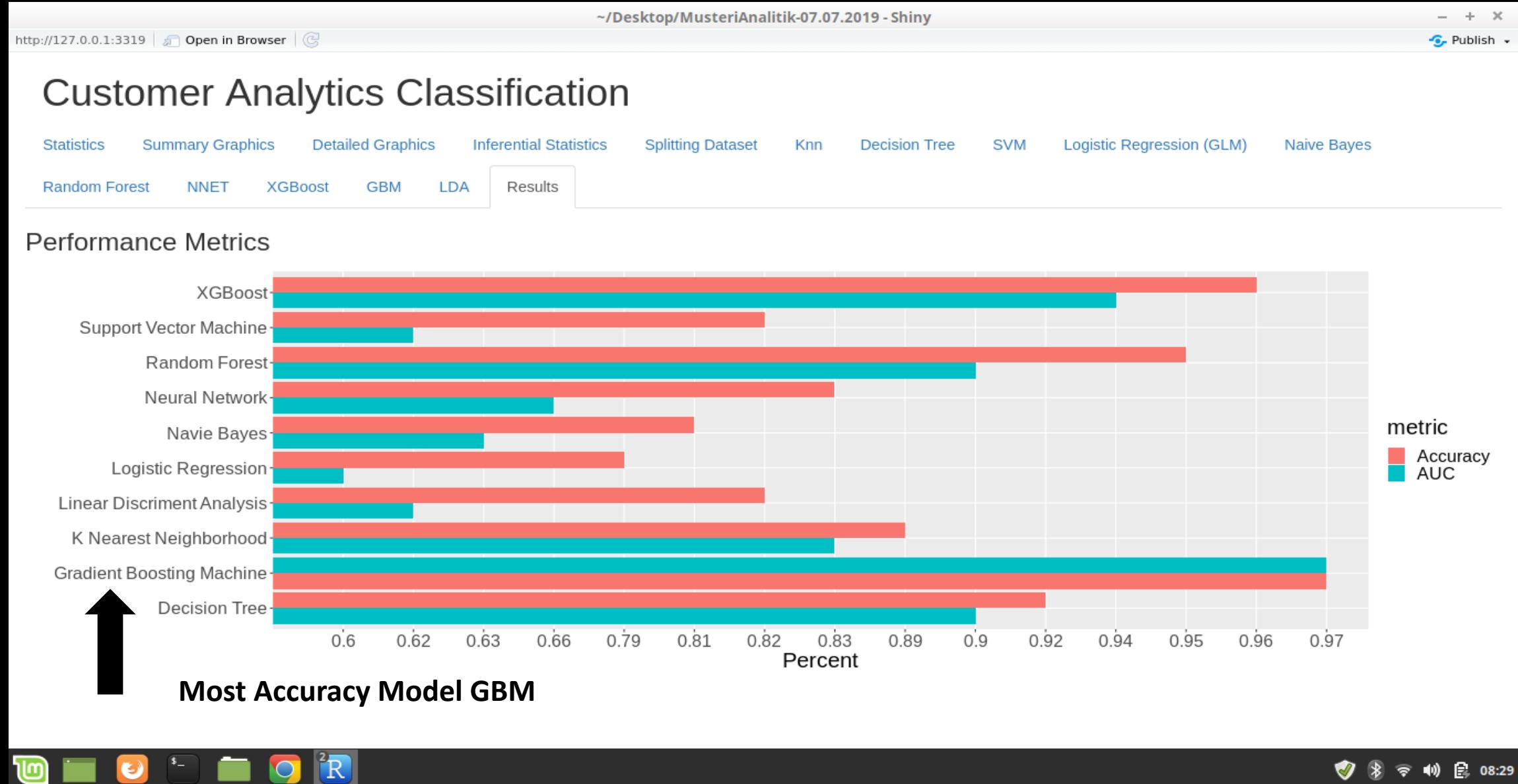
Reference

Prediction	0	1
0	20	2
1	2	82

Accuracy : 0.962  
95% CI : (0.906, 0.99)

08:28

## Result



# Thank You

