CCDS 223 - Data Mining

Final report Universal Bank

Section MB

Dr. Lobna Hsairi

Prepared by

Group (7)

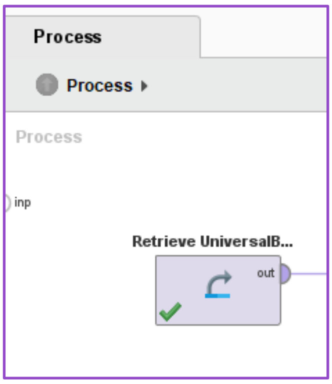
1. **Sara Abdullah (2115103)**
2. **Ahad Abdullah (2111507)**
3. **Sadeem Alzahrani (2111923)**
4. **Muntaha Ali (2110124)**

Universal Bank Case Study

Project: Critical Analysis	
Business Problem Understanding	Universal Bank is a relatively young bank proliferating in terms of overall customer acquisition. The majority of these customers are liability customers (depositors) with varying relationships with the bank. The customer base of asset customers (borrowers) is relatively small, and the bank is interested in expanding this base rapidly to bring in more loan business. In particular, it wants to explore ways of converting its liability customers to **personal loan** customers (while retaining them as depositors). Last year, a campaign the bank ran for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the **Retail Marketing** department to devise more innovative campaigns with better target marketing.
Data Understanding and Collection	The file UniversalBank.xlsx contains data on 5000 customers of Universal Bank (**data courtesy - Statistics.com**). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan offered to them in the earlier campaign. **All data is of type numerical(integer)** **ID (nominal)** Customer ID **Age (Integer)** Customer's age in completed years. **Experience (Integer)** #Of years of professional experience **Income (Integer)** Annual income of the customer ($000) **ZIP Code (Integer)** Home Address ZIP code. **Family (Integer)** Family size of the customer **CCAvg. (Real)** spending on credit cards per month ($000) **Education (Integer)** Education Level: 1: Undergrad; 2: Graduate; 3: Advanced/Professional **Mortgage (Integer)** Value of house mortgage if any. ($000) **Personal Loan (Binary)** Did this customer accept the personal loan offered in the last campaign? **Securities Account (Binary)** Does the customer have a securities account with the bank? **CD Account (Binary)** Does the customer have a certificate of deposit (CD) account with the bank? **Online (Binary)** Does the customer use internet banking facilities? **Credit Card (Binary)** Does the customer use a credit card issued by Universal Bank?

Data Preparation and Feature Selection

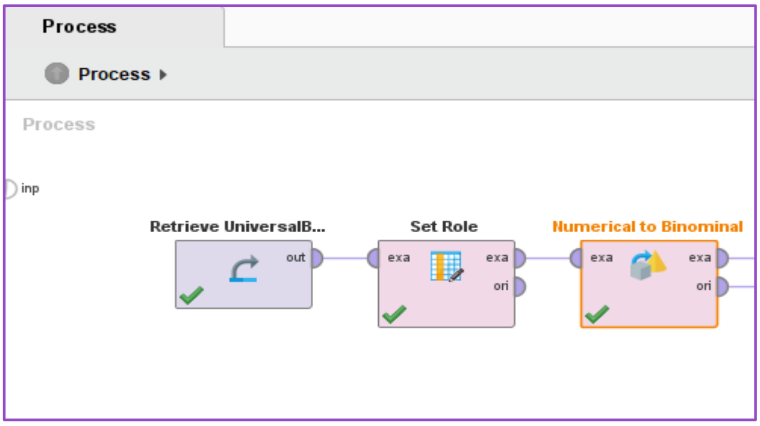
- We load the data in RapidMiner

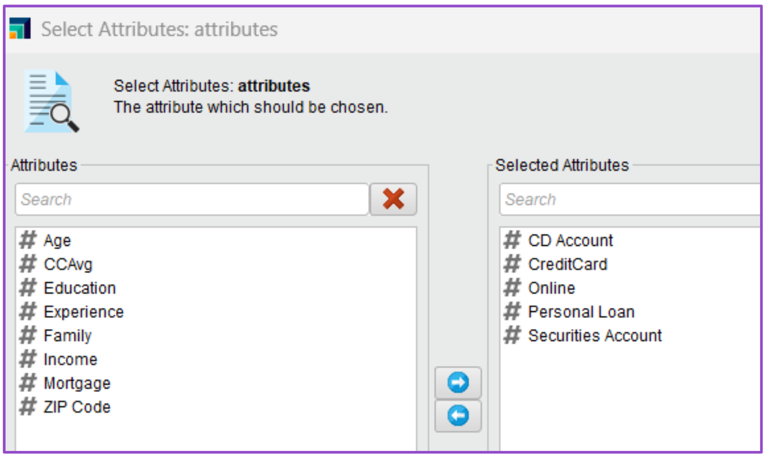


RESULT:

Row No.	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities A...	CD Account	Online	CreditCard
1	25	1	49	91107	4	1.600	1	0	0	1	0	0	0
2	45	19	34	90089	3	1.500	1	0	0	1	0	0	0
3	39	15	11	94720	1	1	1	0	0	0	0	0	0
4	35	9	100	94112	1	2.700	2	0	0	0	0	0	0
5	35	8	45	91330	4	1	2	0	0	0	0	0	1
6	37	13	29	92121	4	0.400	2	155	0	0	0	1	0
7	53	27	72	91711	2	1.500	2	0	0	0	0	1	0
8	50	24	22	93943	1	0.300	3	0	0	0	0	0	1
9	35	10	81	90089	3	0.600	2	104	0	0	0	1	0
10	34	9	180	93023	1	8.900	3	0	1	0	0	0	0
11	65	39	105	94710	4	2.400	3	0	0	0	0	0	0
12	29	5	45	90277	3	0.100	2	0	0	0	0	1	0
13	48	23	114	93106	2	3.800	3	0	0	1	0	0	0
14	59	32	40	94920	4	2.500	2	0	0	0	0	1	0
15	67	41	112	91741	1	2	1	0	0	1	0	0	0
16	60	30	22	95054	1	1.500	3	0	0	0	0	1	1
17	38	14	130	95010	4	4.700	3	134	1	0	0	0	0
18	42	18	81	94305	4	2.400	1	0	0	0	0	0	0
19	46	21	193	91604	2	8.100	3	0	1	0	0	0	0
20	55	28	21	94720	1	0.500	2	0	0	1	0	0	1
21	56	31	25	94015	4	0.900	2	111	0	0	0	1	0
22	57	27	63	90095	3	2	3	0	0	0	0	1	0
23	29	5	62	90277	1	1.200	1	260	0	0	0	1	0
24	44	18	43	91320	2	0.700	1	163	0	1	0	0	0

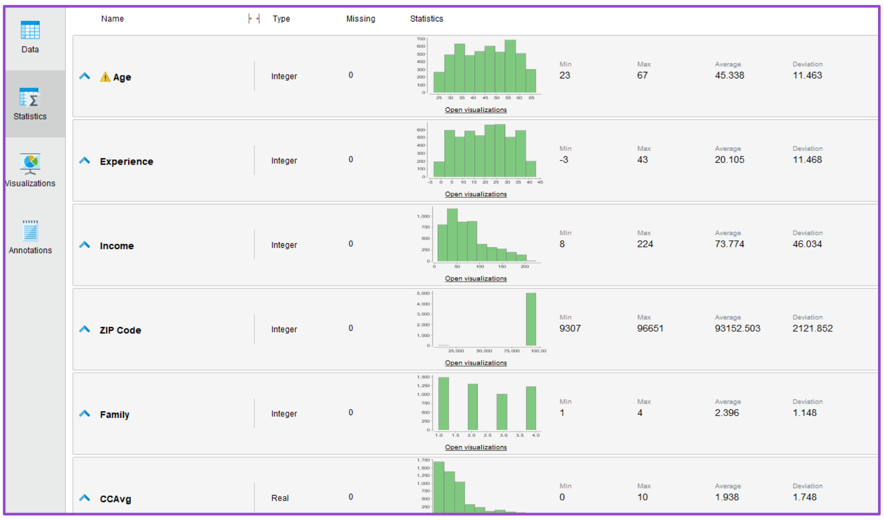
- We dropped the ID column as we already had our Row No. in RapidMiner.
- We can see that all data is in numerical form and binary so we change the binary from 0,1 to true false for better understanding.



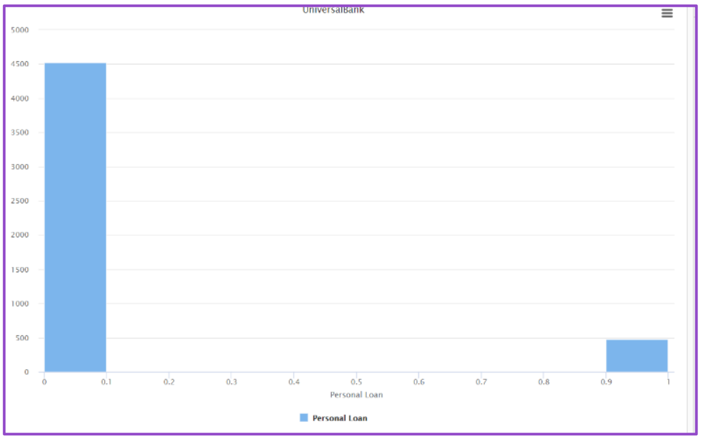


- Also, we Set Role of Personal Loan as label as we have to predict it.

Statistics:



- **We can see the plot of different attributes and we can also visualize data more closely if we want**

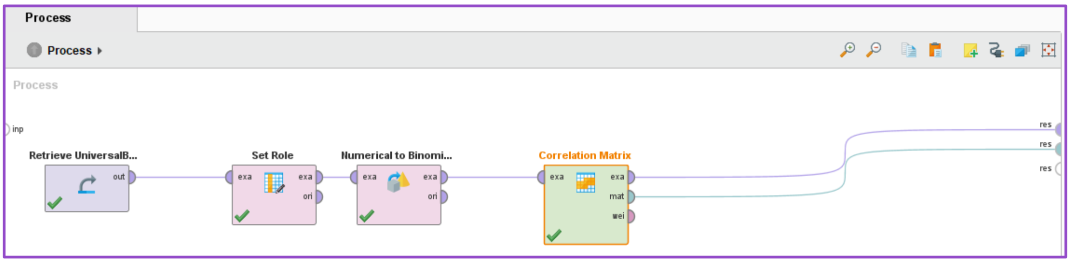


- **We can see that many people did not accept the loan offer.**

Correlation Matrix:

A correlation matrix can help us analyze the relationships between different variables in our dataset. In the context of our scenario, a correlation matrix can provide insights into the potential factors that influence the success of converting liability customers to personal loan customers.

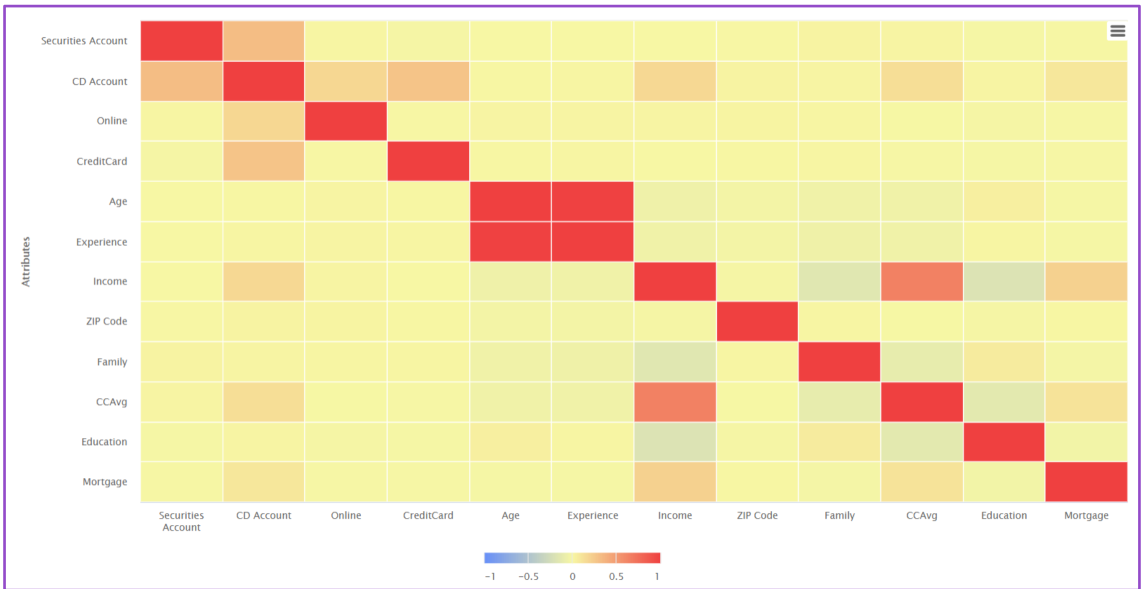
By calculating the correlation coefficients between various variables, we can identify which factors are positively or negatively correlated with the success of the conversion. This information can guide our marketing efforts and help us focus on the most influential factors to improve our campaign's effectiveness.



RESULT:

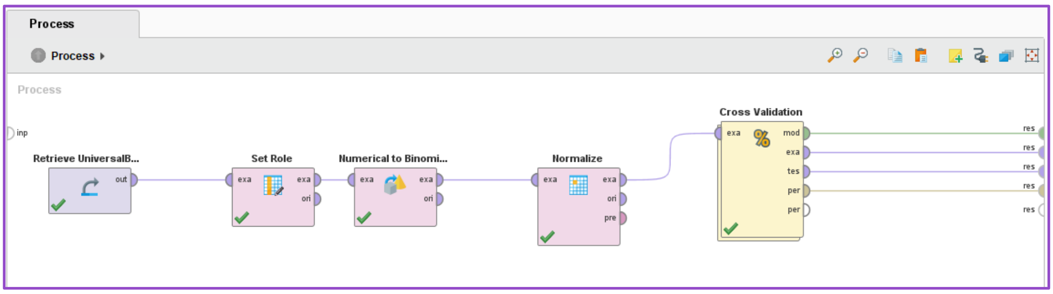
Attributes	Securities Account	CD Account	Online	CreditCard	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage
Securities Account	1	0.317	0.013	-0.015	-0.000	-0.001	-0.003	0.005	0.020	0.015	-0.011	-0.005
CD Account	0.317	1	0.176	0.279	0.008	0.010	0.170	0.020	0.014	0.137	0.014	0.089
Online	0.013	0.176	1	0.004	0.014	0.014	0.014	0.017	0.010	-0.004	-0.015	-0.006
CreditCard	-0.015	0.279	0.004	1	0.008	0.009	-0.002	0.008	0.012	-0.007	-0.011	-0.007
Age	-0.000	0.008	0.014	0.008	1	0.994	-0.055	-0.029	-0.046	-0.052	0.041	-0.013
Experience	-0.001	0.010	0.014	0.009	0.994	1	-0.047	-0.029	-0.053	-0.050	0.013	-0.011
Income	-0.003	0.170	0.014	-0.002	-0.055	-0.047	1	-0.016	-0.158	0.646	-0.188	0.207
ZIP Code	0.005	0.020	0.017	0.008	-0.029	-0.029	-0.016	1	0.012	-0.004	-0.017	0.007
Family	0.020	0.014	0.010	0.012	-0.046	-0.053	-0.158	0.012	1	-0.109	0.065	-0.020
CCAvg	0.015	0.137	-0.004	-0.007	-0.052	-0.050	0.646	-0.004	-0.109	1	-0.136	0.110
Education	-0.011	0.014	-0.015	-0.011	0.041	0.013	-0.188	-0.017	0.065	-0.136	1	-0.033
Mortgage	-0.005	0.089	-0.006	-0.007	-0.013	-0.011	0.207	0.007	-0.020	0.110	-0.033	1

Matrix Visualization:

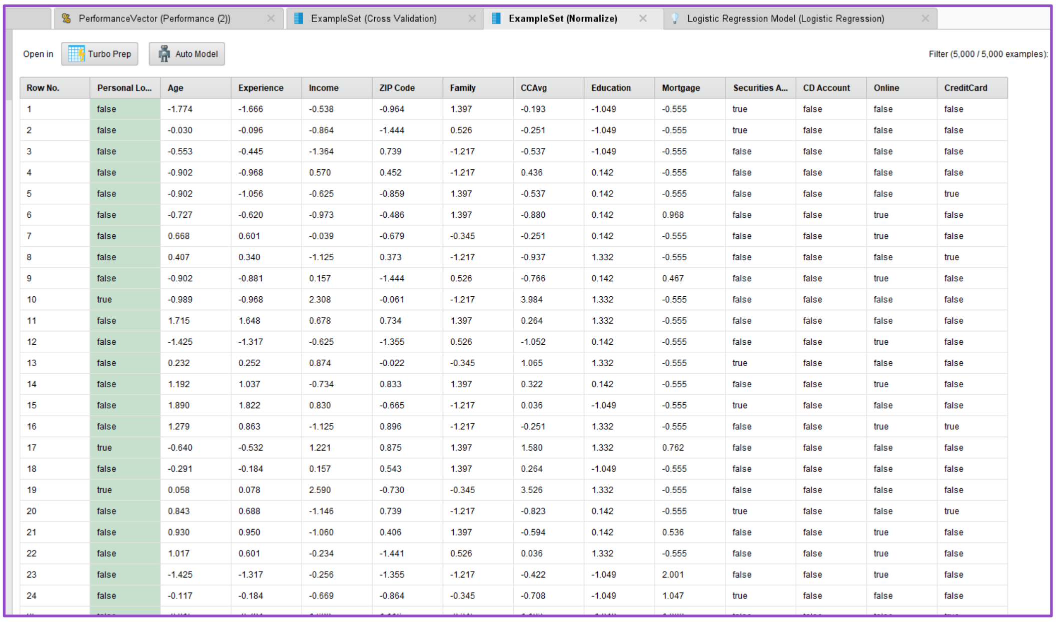


Normalization:

Normalization of data is a preprocessing technique used to scale and transform the features or variables in a dataset. The purpose of normalization is to bring all the variables to a similar scale, ensuring that no variable dominates the others in terms of magnitude. It is particularly important when working with numerical data that has different units, ranges, or distributions.



RESULTS:



PerformanceVector (Performance (2)) | ExampleSet (Cross Validation) | ExampleSet (Normalize) | Logistic Regression Model (Logistic Regression)

Open in Turbo Prep Auto Model

Filter (5,000 / 5,000 examples):

Row No.	Personal Lo...	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Securities A...	CD Account	Online	CreditCard
1	false	-1.774	-1.666	-0.538	-0.964	1.397	-0.193	-1.049	-0.555	true	false	false	false
2	false	-0.030	-0.096	-0.864	-1.444	0.526	-0.251	-1.049	-0.555	true	false	false	false
3	false	-0.553	-0.445	-1.364	0.739	-1.217	-0.537	-1.049	-0.555	false	false	false	false
4	false	-0.902	-0.968	0.570	0.452	-1.217	0.436	0.142	-0.555	false	false	false	false
5	false	-0.902	-1.056	-0.625	-0.859	1.397	-0.537	0.142	-0.555	false	false	false	true
6	false	-0.727	-0.620	-0.973	-0.486	1.397	-0.880	0.142	0.968	false	false	true	false
7	false	0.668	0.601	-0.039	-0.679	-0.345	-0.251	0.142	-0.555	false	false	true	false
8	false	0.407	0.340	-1.125	0.373	-1.217	-0.937	1.332	-0.555	false	false	false	true
9	false	-0.902	-0.881	0.157	-1.444	0.526	-0.766	0.142	0.467	false	false	true	false
10	true	-0.989	-0.968	2.308	-0.061	-1.217	3.984	1.332	-0.555	false	false	false	false
11	false	1.715	1.648	0.678	0.734	1.397	0.264	1.332	-0.555	false	false	false	false
12	false	-1.425	-1.317	-0.625	-1.355	0.526	-1.052	0.142	-0.555	false	false	true	false
13	false	0.232	0.252	0.874	-0.022	-0.345	1.065	1.332	-0.555	true	false	false	false
14	false	1.192	1.037	-0.734	0.833	1.397	0.322	0.142	-0.555	false	false	true	false
15	false	1.890	1.822	0.830	-0.665	-1.217	0.036	-1.049	-0.555	true	false	false	false
16	false	1.279	0.863	-1.125	0.896	-1.217	-0.251	1.332	-0.555	false	false	true	true
17	true	-0.640	-0.532	1.221	0.875	1.397	1.580	1.332	0.762	false	false	false	false
18	false	-0.291	-0.184	0.157	0.543	1.397	0.264	-1.049	-0.555	false	false	false	false
19	true	0.058	0.078	2.590	-0.730	-0.345	3.526	1.332	-0.555	false	false	false	false
20	false	0.843	0.688	-1.146	0.739	-1.217	-0.823	0.142	-0.555	true	false	false	true
21	false	0.930	0.950	-1.060	0.406	1.397	-0.594	0.142	0.536	false	false	true	false
22	false	1.017	0.601	-0.234	-1.441	0.526	0.036	1.332	-0.555	false	false	true	false
23	false	-1.425	-1.317	-0.256	-1.355	-1.217	-0.422	-1.049	2.001	false	false	true	false
24	false	-0.117	-0.184	-0.669	-0.864	-0.345	-0.708	-1.049	1.047	true	false	false	false

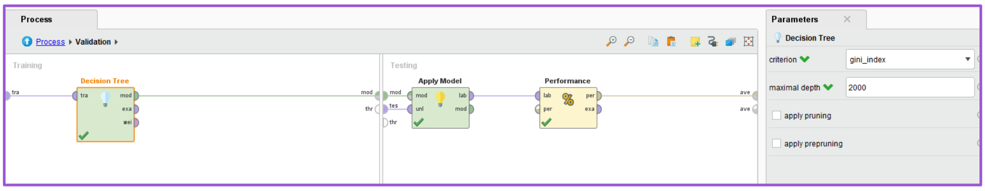
Preprocessing steps such as correlation analysis, normalization, and conversion play crucial roles in preparing data for analysis or modeling tasks. Correlation analysis helps in understanding variable relationships, normalization ensures comparable scales and meets algorithm requirements, and data conversion enables the use of diverse analytical techniques. By performing these preprocessing steps, we enhance the quality, suitability, and interpretability of the data, setting the stage for more accurate and reliable analysis or modeling outcomes.

Modeling Development

- Now after we have prepared our data, we apply decision tree model on our data.
- Now let us split our data into training and testing data using cross validation and apply our model.

Decision tree:

Decision tree with Missing values:



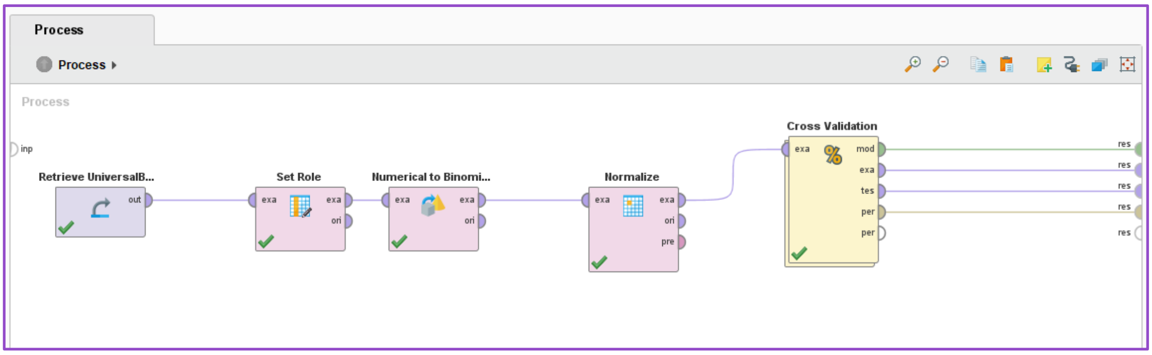
Performance:

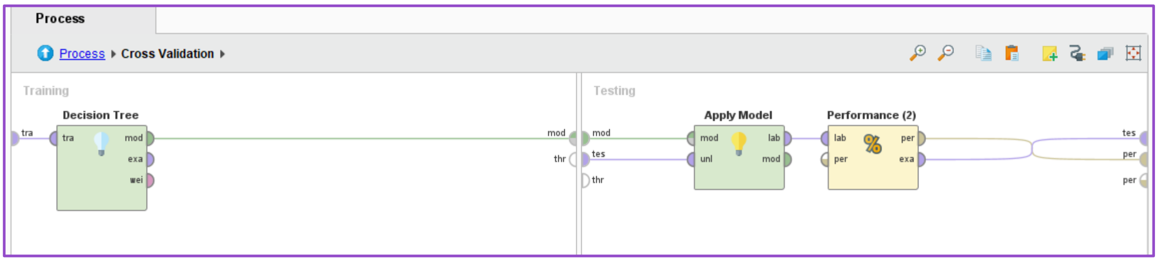
Table View Plot View

accuracy: 88.86%

	true F	true T	class precision
pred. F	462	50	90.23%
pred. T	28	160	85.11%
class recall	94.29%	76.19%	

Decision tree without Missing values:





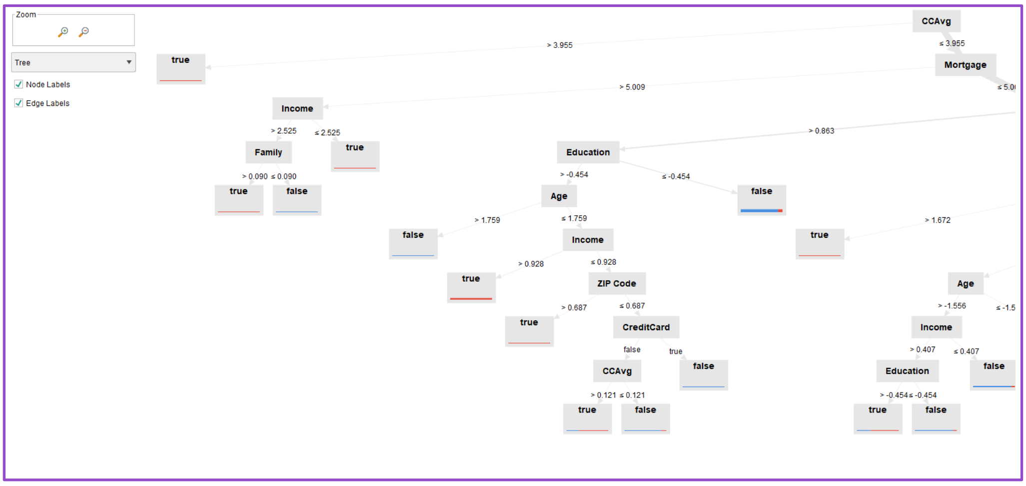
RESULT:

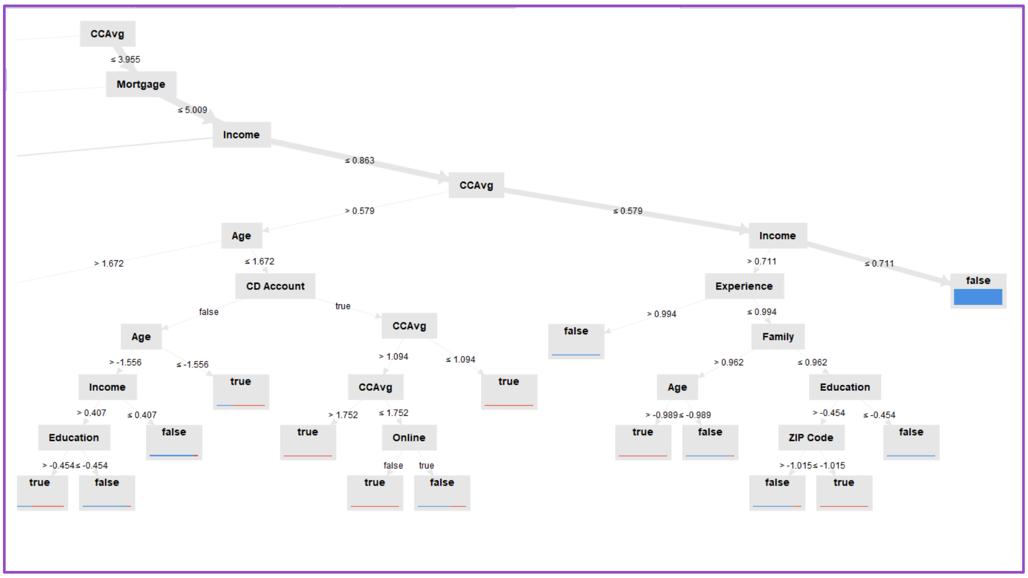
Open in Turbo Prep Auto Model

Filter (5,000 / 5,000 examples): all

Row No.	Personal Lo...	prediction(Personal Loan)	confidence(false)	confidence(true)	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Securities A...	CD A
1	false	false	1	0	-1.774	-1.666	-0.538	-0.964	1.397	-0.193	-1.049	-0.555	true	false
2	false	false	1	0	-0.030	-0.096	-0.864	-1.444	0.526	-0.251	-1.049	-0.555	true	false
3	false	false	1	0	-0.291	-0.184	0.157	0.543	1.397	0.264	-1.049	-0.555	false	false
4	true	true	0	1	-0.640	-0.620	0.982	0.448	-1.217	0.779	0.142	-0.555	false	true
5	false	false	1	0	-1.338	-1.230	-1.212	-0.859	0.526	-0.594	1.332	-0.555	false	false
6	false	false	1	0	-0.553	-0.445	-0.625	1.161	-1.217	-0.708	-1.049	-0.555	false	false
7	false	false	0.900	0.100	0.058	-0.009	0.657	0.430	-1.217	2.153	-1.049	-0.555	false	false
8	false	false	0.894	0.106	-1.251	-1.317	2.481	-0.864	-0.345	1.466	-1.049	3.918	false	false
9	false	false	1	0	1.453	1.386	0.678	1.186	-0.345	0.493	-1.049	2.748	false	false
10	false	false	1	0	0.668	0.776	-1.168	-1.465	1.397	-0.994	-1.049	-0.555	false	false
11	false	false	1	0	0.058	-0.009	-0.973	-0.439	0.526	-0.823	0.142	-0.555	false	false
12	false	false	1	0	-1.600	-1.579	0.765	0.402	1.397	-0.079	1.332	-0.555	false	false
13	false	false	1	0	0.232	0.252	0.005	0.437	-1.217	-0.422	-1.049	-0.555	false	false
14	false	false	1	0	-1.513	-1.404	0.157	0.777	0.526	-0.251	-1.049	2.158	false	false
15	false	false	1	0	-0.378	-0.271	0.200	0.411	1.397	0.417	-1.049	-0.555	false	false
16	false	false	1	0	0.668	0.601	0.396	0.927	-0.345	-0.479	-1.049	-0.555	true	false
17	false	false	1	0	-1.600	-1.666	-0.669	0.732	-1.217	-0.251	-1.049	-0.555	false	false
18	false	false	1	0	-1.076	-0.968	-1.103	0.543	0.526	-0.594	1.332	-0.555	false	false
19	false	false	1	0	0.407	0.514	-1.320	-0.864	1.397	-0.537	-1.049	-0.555	false	false
20	false	false	1	0	-0.815	-0.707	-1.385	0.175	1.397	-0.708	0.142	0.241	false	false
21	false	false	1	0	0.581	0.514	-0.234	-0.205	-0.345	-0.251	0.142	-0.555	true	false
22	false	false	0.894	0.106	0.668	0.776	1.526	-0.215	-0.345	2.782	-1.049	-0.555	false	false
23	false	false	0.894	0.106	0.232	0.340	1.982	0.120	-1.217	1.752	-1.049	-0.555	false	false
24	false	false	1	0	-0.989	-0.881	-0.060	0.454	1.397	-1.052	0.142	-0.555	false	false

TREE:


Zoom
Tree
Node Labels
Edge Labels
CCAvg
> 3.955
≤ 3.955
true
Mortgage
> 5.009
Income
> 2.525
≤ 2.525
Family
true
> 0.090
≤ 0.090
true
false
> 0.863
Education
> -0.454
≤ -0.454
Age
false
> 1.759
≤ 1.759
false
Income
> 0.928
≤ 0.928
true
ZIP Code
> 0.687
≤ 0.687
true
CreditCard
false
true
CCAvg
false
> 0.121
≤ 0.121
true
false
> 1.672
true
Age
> -1.556
Income
> 0.407
≤ 0.407
Education
false
true
false

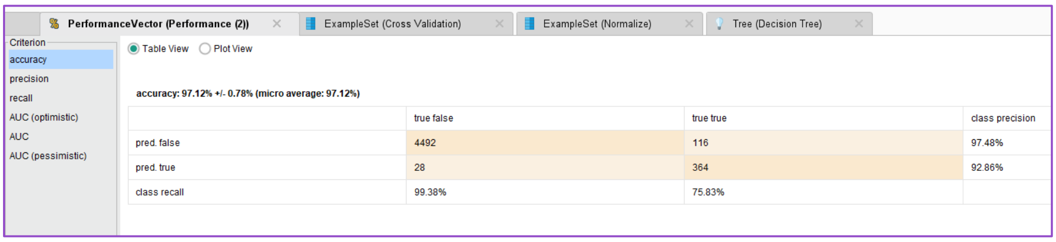

CCAvg
≤ 3.955
Mortgage
≤ 5.009
Income
≤ 0.863
CCAvg
> 0.579
≤ 0.579
Age
> 1.672
≤ 1.672
CD Account
false
true
Age
> -1.556
≤ -1.556
Income
> 0.407
≤ 0.407
Education
> -0.454
≤ -0.454
true
false
false
true
CCAvg
> 1.094
≤ 1.094
CCAvg
> 1.752
≤ 1.752
true
Online
false
true
true
false
true
Income
> 0.711
≤ 0.711
false
Experience
> 0.994
≤ 0.994
false
Family
> 0.962
≤ 0.962
Age
> -0.989
≤ -0.989
true
false
Education
> -0.454
≤ -0.454
ZIP Code
> -1.015
≤ -1.015
false
true
false

Description:

Tree

```
CCAvg > 3.955: true {false=0, true=7}
CCAvg ≤ 3.955
|   Mortgage > 5.009
|   |   Income > 2.525
|   |   |   Family > 0.090: true {false=0, true=2}
|   |   |   Family ≤ 0.090: false {false=2, true=0}
|   |   Income ≤ 2.525: true {false=0, true=9}
|   Mortgage ≤ 5.009
|   |   Income > 0.863
|   |   |   Education > -0.454
|   |   |   |   Age > 1.759: false {false=3, true=0}
|   |   |   |   Age ≤ 1.759
|   |   |   |   |   Income > 0.928: true {false=0, true=298}
|   |   |   |   |   Income ≤ 0.928
|   |   |   |   |   |   ZIP Code > 0.687: true {false=0, true=5}
|   |   |   |   |   |   ZIP Code ≤ 0.687
|   |   |   |   |   |   |   CreditCard = false
|   |   |   |   |   |   |   |   CCAvg > 0.121: true {false=3, true=7}
|   |   |   |   |   |   |   |   CCAvg ≤ 0.121: false {false=7, true=1}
|   |   |   |   |   |   |   CreditCard = true: false {false=4, true=0}
|   |   |   Education ≤ -0.454: false {false=564, true=67}
|   |   Income ≤ 0.863
|   |   |   CCAvg > 0.579
|   |   |   |   Age > 1.672: true {false=0, true=4}
|   |   |   |   Age ≤ 1.672
|   |   |   |   |   CD Account = false
|   |   |   |   |   |   Age > -1.556
|   |   |   |   |   |   |   Income > 0.407
|   |   |   |   |   |   |   |   Education > -0.454: true {false=13, true=26}
|   |   |   |   |   |   |   |   Education ≤ -0.454: false {false=50, true=5}
|   |   |   |   |   |   |   Income ≤ 0.407: false {false=156, true=15}
|   |   |   |   |   |   Age ≤ -1.556: true {false=2, true=4}
|   |   |   |   |   CD Account = true
|   |   |   |   |   |   CCAvg > 1.094
|   |   |   |   |   |   |   CCAvg > 1.752: true {false=0, true=2}
|   |   |   |   |   |   |   CCAvg ≤ 1.752
|   |   |   |   |   |   |   |   Online = false: true {false=0, true=2}
|   |   |   |   |   |   |   |   Online = true: false {false=6, true=3}
|   |   |   |   |   |   CCAvg ≤ 1.094: true {false=0, true=10}
|   |   |   CCAvg ≤ 0.579
|   |   |   |   Income > 0.711
|   |   |   |   |   Experience > 0.994: false {false=20, true=0}
|   |   |   |   |   Experience ≤ 0.994
```

Performance:



PerformanceVector (Performance (2)) | ExampleSet (Cross Validation) | ExampleSet (Normalize) | Tree (Decision Tree)

Criterion: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Table View | Plot View

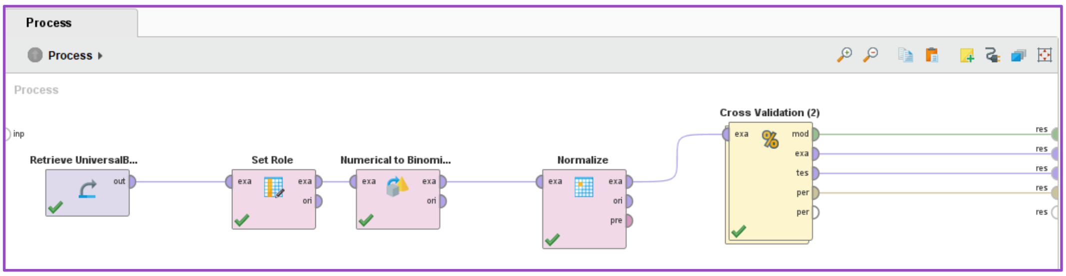
accuracy: 97.12% +/- 0.78% (micro average: 97.12%)

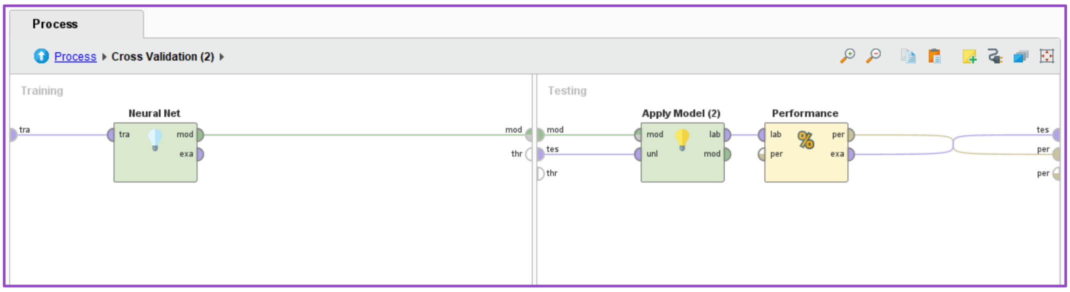
	true false	true true	class precision
pred. false	4492	116	97.48%
pred. true	28	364	92.86%
class recall	99.38%	75.83%	

Changing Parameters:

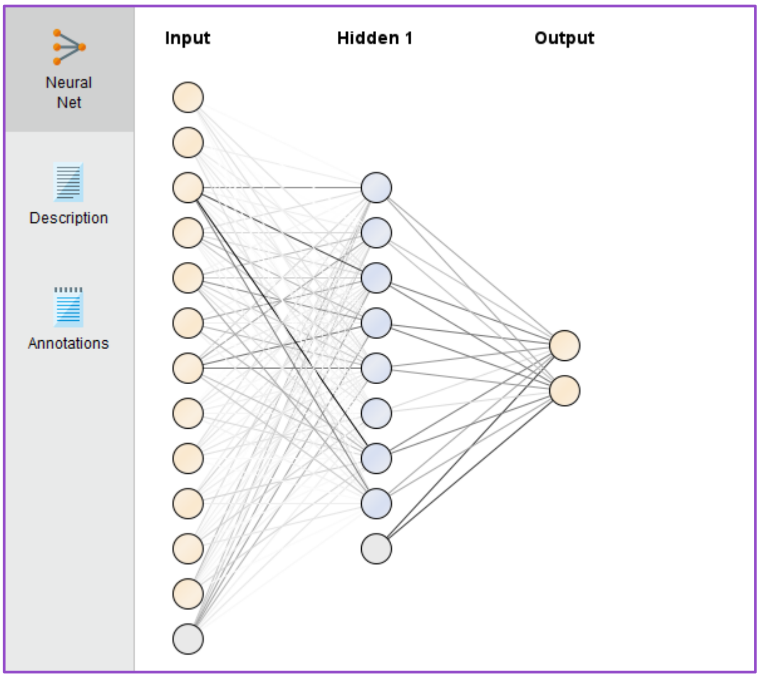
Neural Network with Normalization:

There is not much difference between with and without normalization, so we keep only one.





RESULT:



Open in Turbo Prep | Auto Model

Filter (5,000 / 5,000 examples): all

Row No.	Personal Loan	prediction(Personal Loan)	confidence(false)	confidence(true)	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Securities A...
1	false	false	1.000	0.000	-1.774	-1.666	-0.538	-0.964	1.397	-0.193	-1.049	-0.555	2.929
2	false	false	1.000	0.000	-0.030	-0.096	-0.864	-1.444	0.526	-0.251	-1.049	-0.555	2.929
3	false	false	0.994	0.006	-0.291	-0.184	0.157	0.543	1.397	0.264	-1.049	-0.555	-0.341
4	true	true	0.042	0.958	-0.640	-0.620	0.982	0.448	-1.217	0.779	0.142	-0.555	-0.341
5	false	false	1.000	0.000	-1.338	-1.230	-1.212	-0.859	0.526	-0.594	1.332	-0.555	-0.341
6	false	false	1.000	0.000	-0.553	-0.445	-0.625	1.161	-1.217	-0.708	-1.049	-0.555	-0.341
7	false	false	1.000	0.000	0.058	-0.009	0.657	0.430	-1.217	2.153	-1.049	-0.555	-0.341
8	false	false	0.953	0.047	-1.251	-1.317	2.481	-0.864	-0.345	1.466	-1.049	3.918	-0.341
9	false	false	0.999	0.001	1.453	1.386	0.678	1.186	-0.345	0.493	-1.049	2.748	-0.341
10	false	false	1.000	0.000	0.668	0.776	-1.168	-1.465	1.397	-0.994	-1.049	-0.555	-0.341
11	false	false	1.000	0.000	0.058	-0.009	-0.973	-0.439	0.526	-0.823	0.142	-0.555	-0.341
12	false	false	0.505	0.495	-1.600	-1.579	0.765	0.402	1.397	-0.079	1.332	-0.555	-0.341
13	false	false	1.000	0.000	0.232	0.252	0.005	0.437	-1.217	-0.422	-1.049	-0.555	-0.341
14	false	false	1.000	0.000	-1.513	-1.404	0.157	0.777	0.526	-0.251	-1.049	2.158	-0.341
15	false	false	0.999	0.001	-0.378	-0.271	0.200	0.411	1.397	0.417	-1.049	-0.555	-0.341
16	false	false	1.000	0.000	0.668	0.601	0.396	0.927	-0.345	-0.479	-1.049	-0.555	2.929
17	false	false	1.000	0.000	-1.600	-1.666	-0.669	0.732	-1.217	-0.251	-1.049	-0.555	-0.341
18	false	false	0.999	0.001	-1.076	-0.968	-1.103	0.543	0.526	-0.594	1.332	-0.555	-0.341
19	false	false	1.000	0.000	0.407	0.514	-1.320	-0.864	1.397	-0.537	-1.049	-0.555	-0.341
20	false	false	1.000	0.000	-0.815	-0.707	-1.385	0.175	1.397	-0.708	0.142	0.241	-0.341
21	false	false	1.000	0.000	0.581	0.514	-0.234	-0.205	-0.345	-0.251	0.142	-0.555	2.929
22	false	false	0.988	0.012	0.668	0.776	1.526	-0.215	-0.345	2.782	-1.049	-0.555	-0.341
23	false	false	0.999	0.001	0.232	0.340	1.982	0.120	-1.217	1.752	-1.049	-0.555	-0.341
24	false	false	0.999	0.001	-0.989	-0.881	-0.060	0.454	1.397	-1.052	0.142	-0.555	-0.341

Performance:

PerformanceVector (Performance) | ExampleSet (Cross Validation (2)) | ExampleSet (Normalize) | ImprovedNeuralNet (Neural Net)

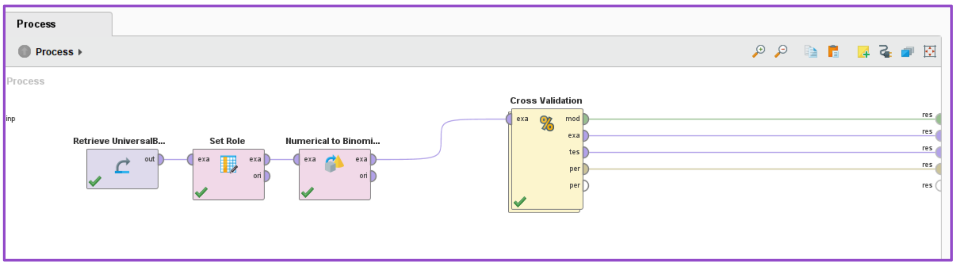
Criterion
- accuracy
- precision
- recall
- AUC (optimistic)
- AUC
- AUC (pessimistic)

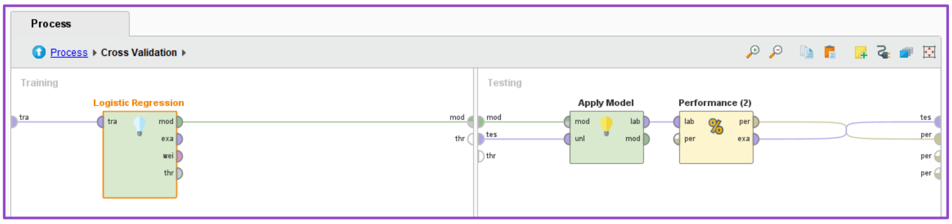
Table View | Plot View

accuracy: 98.02% +/- 0.76% (micro average: 98.02%)

	true false	true true	class precision
pred. false	4484	63	98.61%
pred. true	36	417	92.05%
class recall	99.20%	86.88%	

Logistic Regression:





RESULT:

Open in Turbo Prep Auto Model Filter (5,000 / 5,000 examples): all

Row No.	Personal Loan	prediction(Personal Loan)	confidence(false)	confidence(true)	Securities A...	CD Account	Online	CreditCard	Age	Experience	Income	ZIP Code	Family
1	false	false	0.999	0.001	true	false	false	false	25	1	49	91107	4
2	false	false	1.000	0.000	true	false	false	false	45	19	34	90089	3
3	false	false	0.982	0.018	false	false	false	false	42	18	81	94305	4
4	true	false	0.607	0.393	false	true	true	true	38	13	119	94104	1
5	false	false	0.993	0.007	false	false	false	false	30	6	18	91330	3
6	false	false	1.000	0.000	false	false	true	false	39	15	45	95616	1
7	false	false	0.998	0.002	false	false	true	true	46	20	104	94065	1
8	false	true	0.286	0.714	false	false	false	false	31	5	188	91320	2
9	false	false	0.977	0.023	false	false	false	false	62	36	105	95670	2
10	false	false	1.000	0.000	false	false	true	false	53	29	20	90045	4
11	false	false	0.998	0.002	false	false	false	false	46	20	29	92220	3
12	false	true	0.318	0.682	false	false	false	false	27	2	109	94005	4
13	false	false	0.999	0.001	false	false	true	false	48	23	74	94080	1
14	false	false	0.996	0.004	false	false	true	false	28	4	81	94801	3
15	false	false	0.997	0.003	false	false	true	true	41	17	83	94025	4
16	false	false	0.997	0.003	true	false	false	false	53	27	92	95120	2
17	false	false	1.000	0.000	false	false	true	false	27	1	43	94706	1
18	false	false	0.999	0.001	false	false	true	true	33	9	23	94305	3
19	false	false	1.000	0.000	false	false	true	false	50	26	13	91320	4
20	false	false	0.998	0.002	false	false	false	false	36	12	10	93524	4
21	false	false	0.998	0.002	true	false	true	false	52	26	63	92717	2
22	false	false	0.893	0.107	false	false	true	false	53	29	144	92697	2
23	false	false	0.906	0.094	false	false	false	true	48	24	165	93407	1
24	false	false	0.960	0.040	false	false	false	false	34	10	71	94115	4

Logistic Regression Model

```
Model Metrics Type: BinomialGLM
 Description: N/A
 model id: rm-h2o-model-logistic_regression-11
 frame id: rm-h2o-frame-logistic_regression-11
 MSE: 0.036317896
 RMSE: 0.19057254
 R^2: 0.58151394
 AUC: 0.95884424
 pr_auc: 0.8158653
 logloss: 0.12843426
 mean_per_class_error: 0.16287795
 default threshold: 0.3738589286804199
 CM: Confusion Matrix (Row labels: Actual class; Column labels: Predicted class):
        false  true   Error          Rate
 false   4413   107  0.0237  107 / 4,520
  true    145   335  0.3021    145 / 480
Totals   4558   442  0.0504  252 / 5,000
Gains/Lift Table (Avg response rate:  9.60 %, avg score: 30.66 %):
  Group  Cumulative Data Fraction  Lower Threshold      Lift  Cumulative Lift  Response Rate     Score  Cumulative Response Rate  Cumulative Score  Capture Rate  Cumulative Capture Rate
      1                0.01000000         0.477160  2.500000         2.500000       0.240000  0.552865                  0.240000          0.552865      0.025000                 0.025000
      2                0.02000000         0.453726  1.250000         1.875000       0.120000  0.467532                  0.180000          0.510198      0.012500                 0.037500
      3                0.03000000         0.441851  0.416667         1.388889       0.040000  0.447635                  0.133333          0.489344      0.004167                 0.041667
      4                0.04000000         0.434297  1.250000         1.354167       0.120000  0.438442                  0.130000          0.476618      0.012500                 0.054167
      5                0.05000000         0.426135  1.041667         1.291667       0.100000  0.430039                  0.124000          0.467302      0.010417                 0.064583
      6                0.10000000         0.398676  0.833333         1.062500       0.080000  0.412232                  0.102000          0.439767      0.041667                 0.106250
      7                0.15000000         0.380965  0.791667         0.972222       0.076000  0.388998                  0.093333          0.422844      0.039583                 0.145833
      8                0.20000000         0.365602  1.125000         1.010417       0.108000  0.373167                  0.097000          0.410425      0.056250                 0.202083
      9                0.30000000         0.341429  0.875000         0.965278       0.084000  0.353263                  0.092667          0.391371      0.087500                 0.289583
     10                0.40000000         0.320398  0.833333         0.932292       0.080000  0.331113                  0.089500          0.376306      0.083333                 0.372917
     11                0.50000000         0.301409  0.750000         0.895833       0.072000  0.310881                  0.086000          0.363221      0.075000                 0.447917
     12                0.60000000         0.282741  0.833333         0.885417       0.080000  0.291872                  0.085000          0.351330      0.083333                 0.531250
     13                0.70000000         0.263477  0.729167         0.863095       0.070000  0.273171                  0.082857          0.340164      0.072917                 0.604167
     14                0.80000000         0.243306  1.125000         0.895833       0.108000  0.253535                  0.086000          0.329336      0.112500                 0.716667
     15                0.90000000         0.218102  1.125000         0.921296       0.108000  0.231501                  0.088444          0.318465      0.112500                 0.829167
     16                1.00000000         0.145002  1.708333         1.000000       0.164000  0.199973                  0.096000          0.306616      0.170833                 1.000000
 null DOF: 4999.0
 residual DOF: 4987.0
 null deviance: 3162.041
 residual deviance: 1284.3425
GLM Model (summary):
   Family  Link Regularization Number of Predictors Total Number of Active Predictors Number of Iterations                 Training Frame
 binomial logit           None                   12                                12                    7 rm-h2o-frame-logistic_regression-11
```

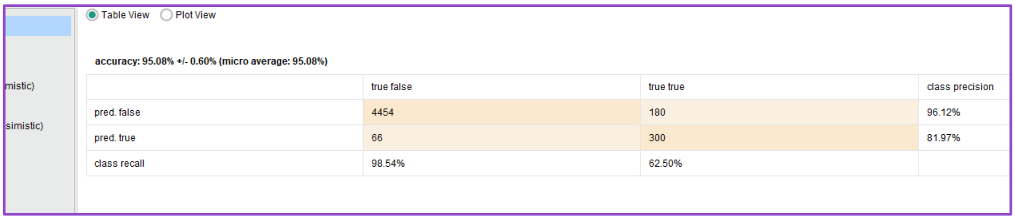


Table View Plot View

mistic)

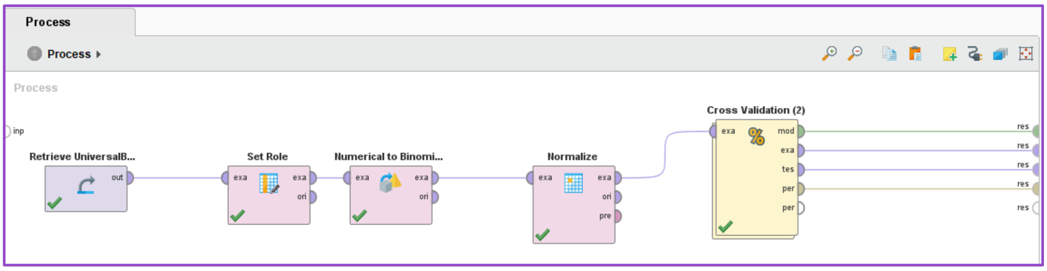
simistic)

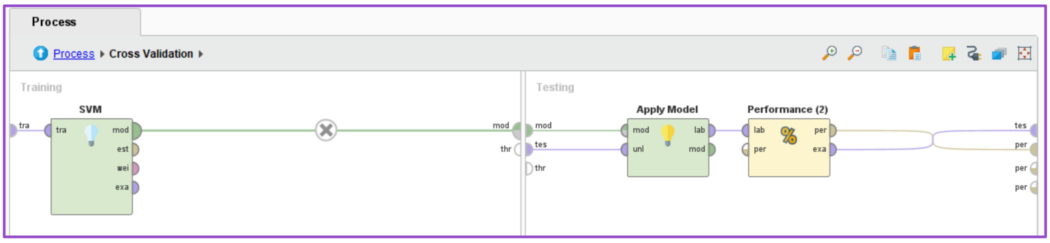
accuracy: 95.08% +/- 0.60% (micro average: 95.08%)

	true false	true true	class precision
pred. false	4454	180	96.12%
pred. true	66	300	81.97%
class recall	98.54%	62.50%	

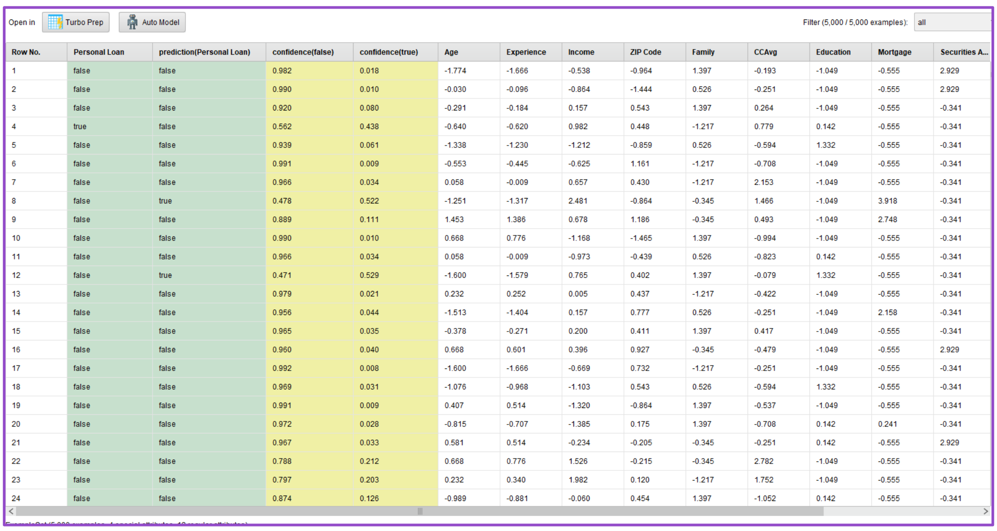
- We applied model and predicted the value whether loan offer will be accepted or not and based on the given data we designed the result and also checked the performance of our data.

SVM:



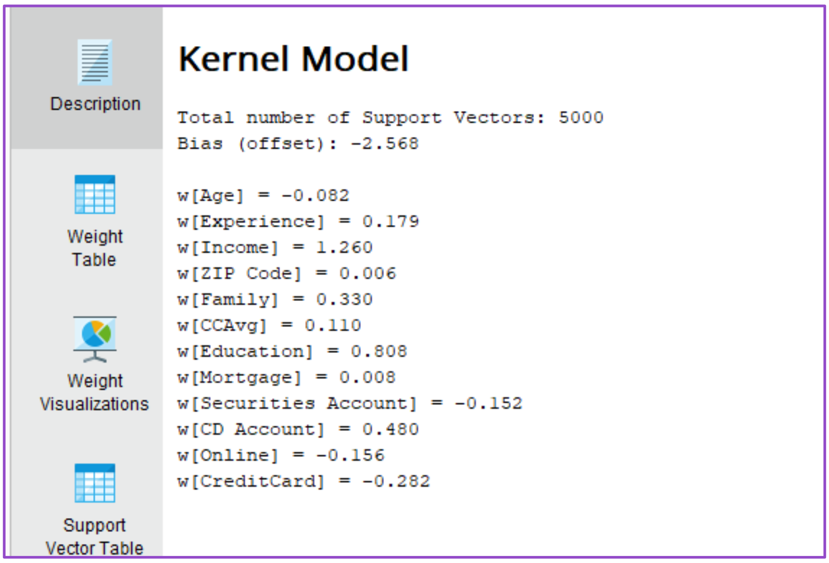


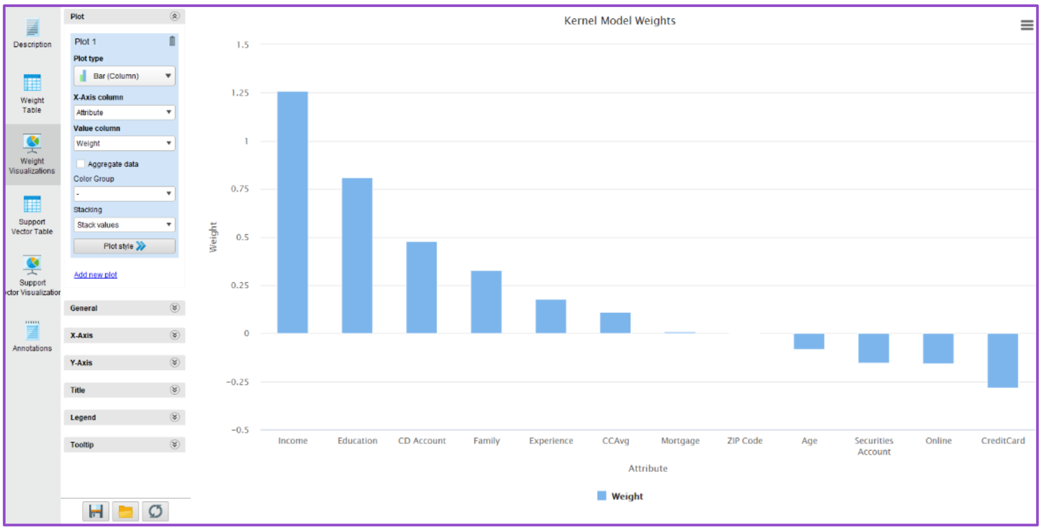
RESULT:



Open in Turbo Prep Auto Model Filter (5,000 / 5,000 examples): all

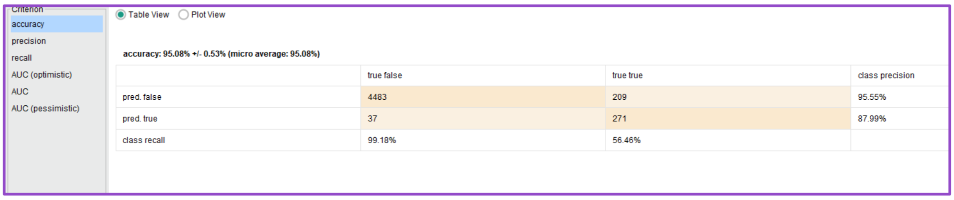
Row No.	Personal Loan	prediction(Personal Loan)	confidence(false)	confidence(true)	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Securities A...
1	false	false	0.982	0.018	-1.774	-1.666	-0.538	-0.964	1.397	-0.193	-1.049	-0.555	2.929
2	false	false	0.990	0.010	-0.030	-0.096	-0.864	-1.444	0.526	-0.251	-1.049	-0.555	2.929
3	false	false	0.920	0.080	-0.291	-0.184	0.157	0.543	1.397	0.264	-1.049	-0.555	-0.341
4	true	false	0.562	0.438	-0.640	-0.620	0.982	0.448	-1.217	0.779	0.142	-0.555	-0.341
5	false	false	0.939	0.061	-1.338	-1.230	-1.212	-0.859	0.526	-0.594	1.332	-0.555	-0.341
6	false	false	0.991	0.009	-0.553	-0.445	-0.625	1.161	-1.217	-0.708	-1.049	-0.555	-0.341
7	false	false	0.966	0.034	0.058	-0.009	0.657	0.430	-1.217	2.153	-1.049	-0.555	-0.341
8	false	true	0.478	0.522	-1.251	-1.317	2.481	-0.864	-0.345	1.466	-1.049	3.918	-0.341
9	false	false	0.889	0.111	1.453	1.386	0.678	1.186	-0.345	0.493	-1.049	2.748	-0.341
10	false	false	0.990	0.010	0.668	0.776	-1.168	-1.465	1.397	-0.994	-1.049	-0.555	-0.341
11	false	false	0.966	0.034	0.058	-0.009	-0.973	-0.439	0.526	-0.823	0.142	-0.555	-0.341
12	false	true	0.471	0.529	-1.600	-1.579	0.765	0.402	1.397	-0.079	1.332	-0.555	-0.341
13	false	false	0.979	0.021	0.232	0.252	0.005	0.437	-1.217	-0.422	-1.049	-0.555	-0.341
14	false	false	0.956	0.044	-1.513	-1.404	0.157	0.777	0.526	-0.251	-1.049	2.158	-0.341
15	false	false	0.965	0.035	-0.378	-0.271	0.200	0.411	1.397	0.417	-1.049	-0.555	-0.341
16	false	false	0.960	0.040	0.668	0.601	0.396	0.927	-0.345	-0.479	-1.049	-0.555	2.929
17	false	false	0.992	0.008	-1.600	-1.666	-0.669	0.732	-1.217	-0.251	-1.049	-0.555	-0.341
18	false	false	0.969	0.031	-1.076	-0.968	-1.103	0.543	0.526	-0.594	1.332	-0.555	-0.341
19	false	false	0.991	0.009	0.407	0.514	-1.320	-0.864	1.397	-0.537	-1.049	-0.555	-0.341
20	false	false	0.972	0.028	-0.815	-0.707	-1.385	0.175	1.397	-0.708	0.142	0.241	-0.341
21	false	false	0.967	0.033	0.581	0.514	-0.234	-0.205	-0.345	-0.251	0.142	-0.555	2.929
22	false	false	0.788	0.212	0.668	0.776	1.526	-0.215	-0.345	2.782	-1.049	-0.555	-0.341
23	false	false	0.797	0.203	0.232	0.340	1.982	0.120	-1.217	1.752	-1.049	-0.555	-0.341
24	false	false	0.874	0.126	-0.989	-0.881	-0.060	0.454	1.397	-1.052	0.142	-0.555	-0.341


Description
Weight Table
Weight Visualizations
Support Vector Table
Kernel Model
Total number of Support Vectors: 5000
Bias (offset): -2.568
w[Age] = -0.082
w[Experience] = 0.179
w[Income] = 1.260
w[ZIP Code] = 0.006
w[Family] = 0.330
w[CCAvg] = 0.110
w[Education] = 0.808
w[Mortgage] = 0.008
w[Securities Account] = -0.152
w[CD Account] = 0.480
w[Online] = -0.156
w[CreditCard] = -0.282


Kernel Model Weights
Weight
Income
Education
CD Account
Family
Experience
CCAvg
Mortgage
ZIP Code
Age
Securities Account
Online
CreditCard
Attribute
Weight


Result History
PerformanceVector (Performance (2))
ExampleSet (Cross Validation)
ExampleSet (Normalize)
Kernel Model (SVM)
Function Value
Counter
false
true

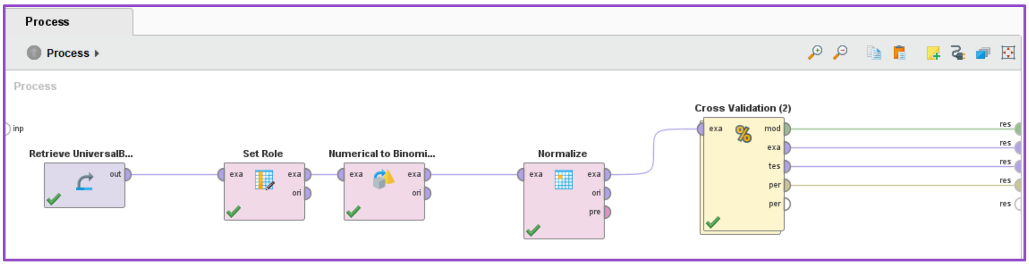
Performance:

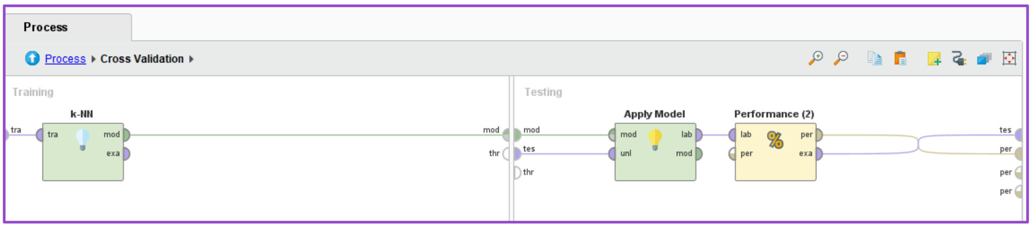


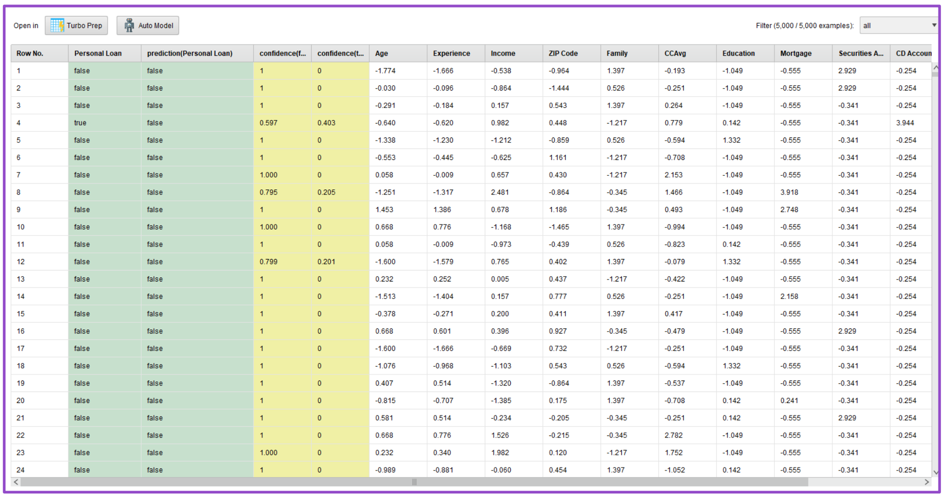
accuracy: 95.08% +/- 0.53% (micro average: 95.08%)

	true false	true true	class precision
pred. false	4483	209	95.55%
pred. true	37	271	87.99%
class recall	99.18%	56.46%	

K-NN:





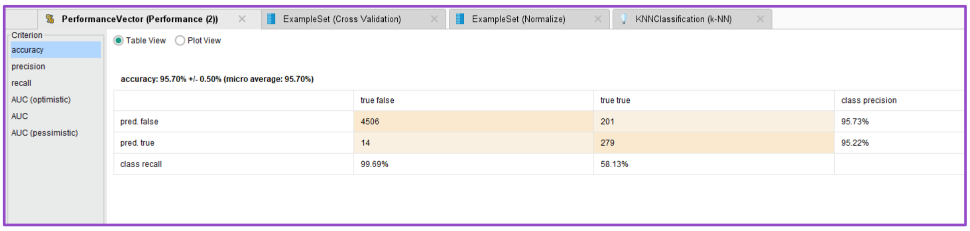


Open in Turbo Prep Auto Model

Filter (5,000 / 5,000 examples): all

Row No.	Personal Loan	prediction(Personal Loan)	confidence(f...	confidence(t...	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Securities A...	CD Accoun
1	false	false	1	0	-1.774	-1.666	-0.538	-0.964	1.397	-0.193	-1.049	-0.555	2.929	-0.254
2	false	false	1	0	-0.030	-0.096	-0.864	-1.444	0.526	-0.251	-1.049	-0.555	2.929	-0.254
3	false	false	1	0	-0.291	-0.184	0.157	0.543	1.397	0.264	-1.049	-0.555	-0.341	-0.254
4	true	false	0.597	0.403	-0.640	-0.620	0.982	0.448	-1.217	0.779	0.142	-0.555	-0.341	3.944
5	false	false	1	0	-1.338	-1.230	-1.212	-0.859	0.526	-0.594	1.332	-0.555	-0.341	-0.254
6	false	false	1	0	-0.553	-0.445	-0.625	1.161	-1.217	-0.708	-1.049	-0.555	-0.341	-0.254
7	false	false	1.000	0	0.058	-0.009	0.657	0.430	-1.217	2.153	-1.049	-0.555	-0.341	-0.254
8	false	false	0.795	0.205	-1.251	-1.317	2.481	-0.864	-0.345	1.466	-1.049	3.918	-0.341	-0.254
9	false	false	1	0	1.453	1.386	0.678	1.186	-0.345	0.493	-1.049	2.748	-0.341	-0.254
10	false	false	1.000	0	0.668	0.776	-1.168	-1.465	1.397	-0.994	-1.049	-0.555	-0.341	-0.254
11	false	false	1	0	0.058	-0.009	-0.973	-0.439	0.526	-0.823	0.142	-0.555	-0.341	-0.254
12	false	false	0.799	0.201	-1.600	-1.579	0.765	0.402	1.397	-0.079	1.332	-0.555	-0.341	-0.254
13	false	false	1	0	0.232	0.252	0.005	0.437	-1.217	-0.422	-1.049	-0.555	-0.341	-0.254
14	false	false	1	0	-1.513	-1.404	0.157	0.777	0.526	-0.251	-1.049	2.158	-0.341	-0.254
15	false	false	1	0	-0.378	-0.271	0.200	0.411	1.397	0.417	-1.049	-0.555	-0.341	-0.254
16	false	false	1	0	0.668	0.601	0.396	0.927	-0.345	-0.479	-1.049	-0.555	2.929	-0.254
17	false	false	1	0	-1.600	-1.666	-0.669	0.732	-1.217	-0.251	-1.049	-0.555	-0.341	-0.254
18	false	false	1	0	-1.076	-0.968	-1.103	0.543	0.526	-0.594	1.332	-0.555	-0.341	-0.254
19	false	false	1	0	0.407	0.514	-1.320	-0.864	1.397	-0.537	-1.049	-0.555	-0.341	-0.254
20	false	false	1	0	-0.815	-0.707	-1.385	0.175	1.397	-0.708	0.142	0.241	-0.341	-0.254
21	false	false	1	0	0.581	0.514	-0.234	-0.205	-0.345	-0.251	0.142	-0.555	2.929	-0.254
22	false	false	1	0	0.668	0.776	1.526	-0.215	-0.345	2.782	-1.049	-0.555	-0.341	-0.254
23	false	false	1.000	0	0.232	0.340	1.982	0.120	-1.217	1.752	-1.049	-0.555	-0.341	-0.254
24	false	false	1	0	-0.989	-0.881	-0.060	0.454	1.397	-1.052	0.142	-0.555	-0.341	-0.254

Performance:



PerformanceVector (Performance (2)) | ExampleSet (Cross Validation) | ExampleSet (Normalize) | KNNClassification (k-NN)

accuracy: 95.70% +/- 0.50% (micro average: 95.70%)

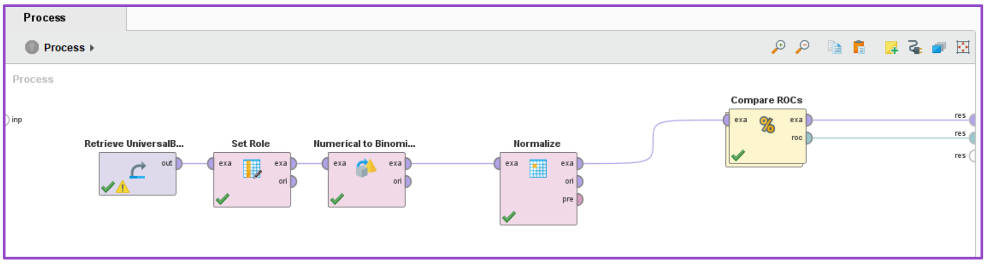
	true false	true true	class precision
pred. false	4506	201	95.73%
pred. true	14	279	95.22%
class recall	99.69%	58.13%	

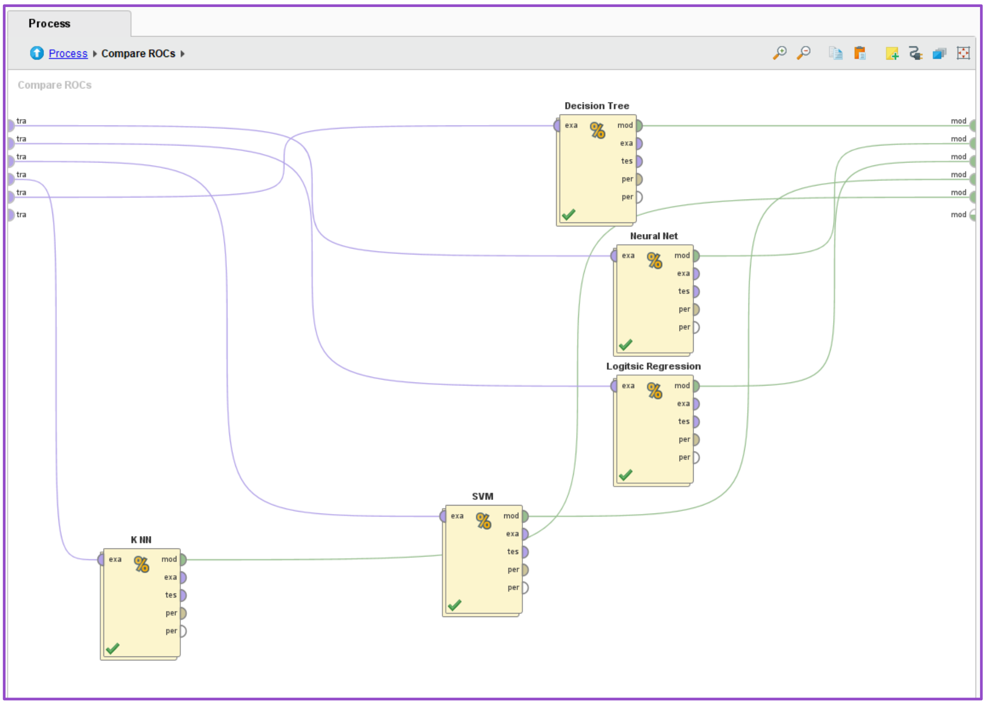
ROC Comparison:

When comparing different classifiers using Receiver Operating Characteristic (ROC) analysis, we found that the k-nearest neighbors (KNN) classifier performed the best. The ROC analysis is a popular method for evaluating and comparing the performance of classification models, particularly when dealing with imbalanced datasets or scenarios where the trade-off between true positives and false positives is crucial.

- ROC analysis: ROC analysis provides a comprehensive assessment of a classifier's performance by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds. The resulting ROC curve illustrates the trade-off between sensitivity and specificity for different threshold settings.
- Comparison of classifiers: By using ROC analysis, you compared the performance of multiple classifiers on your dataset. Each classifier generates its own ROC curve, allowing you to assess how well they distinguish between positive and negative instances.
- KNN as the best classifier: In your analysis, you found that the KNN classifier outperformed the other classifiers based on the ROC curves. This means that the KNN classifier achieved a higher true positive rate (sensitivity) while maintaining a lower false positive rate (1 - specificity) compared to the other classifiers at different threshold values.
- Interpretation of results: The fact that KNN performed the best according to the ROC analysis suggests that it was more effective in correctly classifying positive instances (loan acceptance) while minimizing the misclassification of negative instances (loan rejection). This indicates that KNN captured the underlying patterns and relationships in the data well, enabling it to make more accurate predictions compared to the other classifiers you tested.
- Practical implications: The finding that KNN was the best classifier in terms of ROC analysis has practical implications for your project. It suggests that KNN is a suitable choice for predicting loan acceptance based on the attributes and factors you considered. Financial institutions can potentially leverage the KNN classifier to evaluate loan applications and make more accurate predictions about the likelihood of acceptance.

It's important to note that while ROC analysis provides valuable insights into classifier performance, it doesn't provide the complete picture. Other evaluation metrics such as accuracy, precision, recall, or F1 score should also be considered to assess the overall effectiveness of the KNN classifier and validate its superiority over other classifiers.

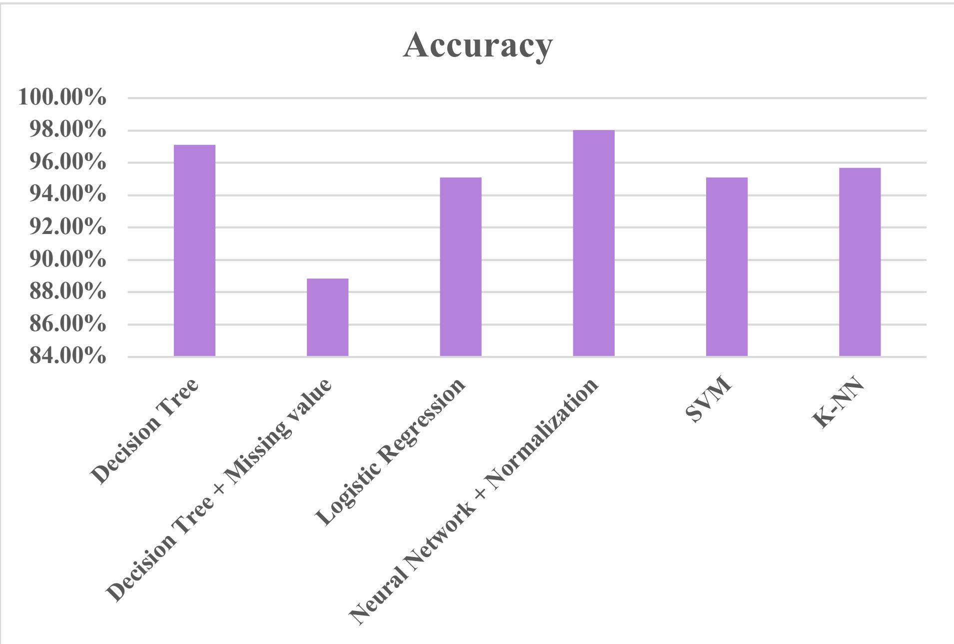

Process
Process
Retrieve UniversalB...
Set Role
Numerical to Binomi...
Normalize
Compare ROCs


Process
Process
Compare ROCs
Decision Tree
Neural Net
Logitsic Regression
SVM
K NN

RESULT:


ROC Comparison (Compare ROCs)
ExampleSet (Normalize)
k-NN
Logistic Regression (6)
SVM (2)
Decision Tree (2)
Neural Net (2)

Experiment	Accuracy
Decision Tree	97.12 %
Decision Tree + Missing value	88.86 %
Logistic Regression	95.08 %
Neural Network + Normalization	98.02 %
SVM	95.08 %
K-NN	95.70 %



Model Deployment (Assess and Reflection):

The project y aims to study the factors that influence the acceptance of a loan offer and explore the use of both manual and automated methods to predict loan acceptance. By designing a model that incorporates various attributes, you can analyze how these factors impact the likelihood of a loan offer being accepted.

- **Factors affecting loan acceptance**: The project focuses on identifying the specific attributes or variables that have a significant influence on whether a loan offer is accepted or not. These factors can include a wide range of variables such as income, credit history, employment status, loan amount, interest rate, loan term, and other relevant financial and personal information.

- **Manual and automated methods**: The project explores the use of both manual and automated methods to calculate and predict loan acceptance. Manual methods may involve traditional underwriting processes where human experts analyze the applicant's information and make a subjective judgment. Automated methods, on the other hand, leverage machine learning algorithms, such as decision trees or other predictive models, to analyze the data and make predictions based on learned patterns and relationships.

- **Model design and comparison**: In this project, you design a model that incorporates the identified attributes and factors influencing loan acceptance. This model serves as a framework to predict whether a loan offer will be accepted or not. By comparing the results obtained from different attributes and factors, you can gain insights into which variables have the most significant impact on loan acceptance.

- **Role in the banking and finance sector:** The project's findings and the developed model can play a crucial role in the banking and finance sector. By analyzing and predicting loan acceptance based on various factors, financial institutions can make informed decisions on loan approvals and manage risk effectively. This project can contribute to streamlining the lending process, improving efficiency, and minimizing the chances of defaults.

- **Predictive capabilities**: The project's ultimate goal is to use the developed model to predict loan acceptance for different individuals based on their specific details and attributes. By inputting an applicant's information into the model, financial institutions can obtain a prediction of the likelihood of that particular loan offer being accepted. This prediction can assist in making informed decisions and improving loan approval rates.

Overall, the project focuses on understanding the factors influencing loan acceptance, exploring manual and automated methods to calculate loan acceptance, designing a model based on these factors, and comparing the results obtained. Its outcomes can have significant implications for the banking and finance sector by aiding in the prediction of loan acceptance and facilitating more informed decision-making processes.