**Project Analytics - DSTI - 7/7/2024**

**Phuc Nguyen PHAM**

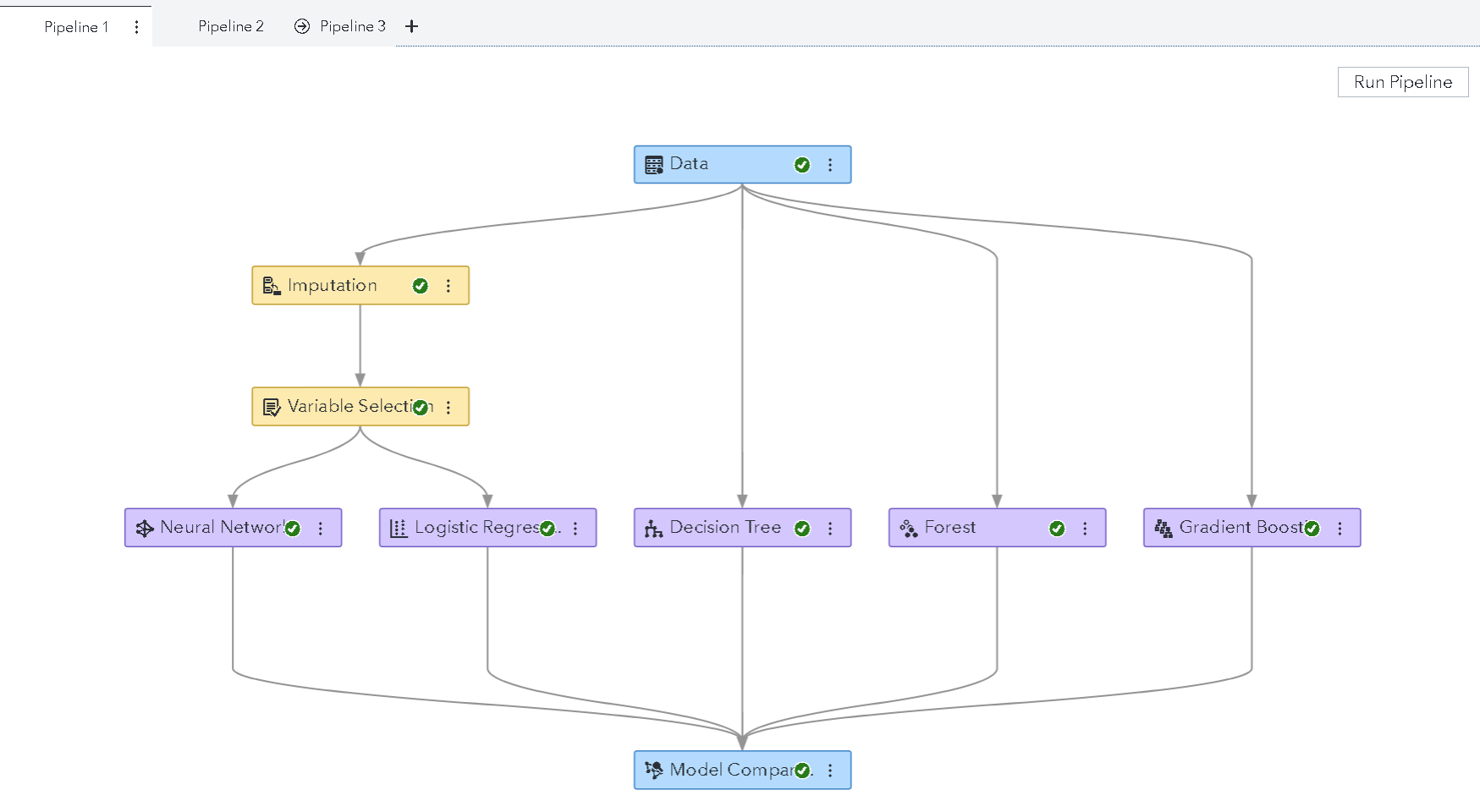
Data Mining & Machine Learning on SAS Viya – BigOrganics Business Case –

First of all, we setup the parameters so they are ready to be trained. 2 variables “DemCluster” and “TargetAmt” will be rejected.

A screenshot of a computer

Description automatically generated

We tried the first data model like this:



The results show that the Forest gives the highest score with (0.6334).

A screenshot of a computer

Description automatically generated

With the second data model, the result is still the same.

A diagram of a company

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

With the third data model, we can see that the forest give a better score (0.6548)

A screenshot of a computer

Description automatically generated

A white background with black text

Description automatically generated

So we conclude that the third pipeline is the best. We did 2 transformation, 1 imputation, 1 ensemble and 1 feature machine. Those help to make a better result.

Ex 2:

First of all, We changed Target Gift Flag as a Category.

A screenshot of a computer

Description automatically generated

Here is the list of predictors that we choose:

A list of black and white check marks

Description automatically generated

Then we create some models.

1. We got the result of 0.1639 (Youden) for the decision tree.

A screenshot of a computer

Description automatically generated

1. For the Bayesian Network, we got 0.499

A screenshot of a computer

Description automatically generated

1. For the Gradient Boosting, we got 0.3488.

A screenshot of a computer

Description automatically generated

1. For the Forest model, we got 0.1563

A screenshot of a computer

Description automatically generated

1. We got the score of 0.1863 for the Logistic Regression.

A screenshot of a computer

Description automatically generated

1. For the Neural Network, we got the score of 0.

A screenshot of a computer

Description automatically generated

1. And finally, we got the score of 0.1783 for the SVM.

A screenshot of a computer

Description automatically generated

Then we build a data comparison tab:

A screenshot of a computer

Description automatically generated

According to the model comparison, Bayesian network is the best model with the Youden score at 0.449, so it will be selected.

We export to the Excel file, then go to the tab “Lift” and create a new column with the formula like this:

=100000\*A2/100\*(-2+3\*F2\*50/100)

We choose the formula like it because there is more than 100 000 people in the PVA Table, and there is 50% of people who are likely to do the donation. The cost is 2$. The margin is 3$.

So we can get the result like this:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Percentile | Observations | Events | Model | Best | Cumulative Model | Cumulative Best | ROI |
| 5.00 | 3250 | 3250 | 1.9055 | 1.9055 | 1.9055 | 1.9055 | 4291.577497 |
| 10.00 | 3250 | 3173 | 1.8604 | 1.9055 | 1.8830 | 1.9055 | 8244.554543 |
| 15.00 | 3250 | 2821 | 1.6540 | 1.9055 | 1.8067 | 1.9055 | 10649.64381 |
| 20.00 | 3250 | 2755 | 1.6153 | 1.9055 | 1.7588 | 1.9055 | 12764.50412 |
| 25.00 | 3250 | 2557 | 1.4992 | 1.9055 | 1.7069 | 1.9055 | 14008.67755 |
| 30.00 | 3250 | 2282 | 1.3380 | 1.9055 | 1.6454 | 1.9055 | 14043.56366 |
| 35.00 | 3250 | 2095 | 1.2283 | 1.9055 | 1.5858 | 1.9055 | 13256.13438 |
| 40.00 | 3250 | 1879 | 1.1017 | 1.9055 | 1.5253 | 1.9055 | 11518.86488 |
| 45.00 | 3250 | 1816 | 1.0648 | 1.9055 | 1.4741 | 1.9055 | 9504.558647 |
| 50.00 | 3250 | 1611 | 0.9446 | 1.9055 | 1.4212 | 1.9055 | 6588.783677 |
| 55.00 | 3250 | 1589 | 0.9317 | 0.9446 | 1.3767 | 1.8182 | 3576.265721 |
| 60.00 | 3250 | 1639 | 0.9610 | 0.0000 | 1.3420 | 1.6667 | 783.6181877 |
| 65.00 | 3250 | 1485 | 0.8707 | 0.0000 | 1.3058 | 1.5385 | -2686.23025 |
| 70.00 | 3250 | 1280 | 0.7505 | 0.0000 | 1.2661 | 1.4286 | -7057.54742 |
| 75.00 | 3250 | 1149 | 0.6737 | 0.0000 | 1.2266 | 1.3333 | -12004.9251 |
| 80.00 | 3250 | 1113 | 0.6526 | 0.0000 | 1.1907 | 1.2500 | -17110.6095 |
| 85.00 | 3250 | 891 | 0.5224 | 0.0000 | 1.1514 | 1.1765 | -23192.5185 |
| 90.00 | 3250 | 451 | 0.2644 | 0.0000 | 1.1022 | 1.1111 | -31209.2873 |
| 95.00 | 3250 | 275 | 0.1612 | 0.0000 | 1.0526 | 1.0526 | -40000 |
| 100.00 | 3238 | 0 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | -50000 |

Then we visualize the result:

According to the ROI graph, the maximum of the ROI (which is 14043.56$) is at the percentile of 30%. It means that if we select the 30% who has the highest probability to donate, we must get 14043.56$ back.

Please find the attached to see the project.