Final Project Report

IST 707

Daniel Piston

2021

Contents

[Introduction 2](#_Toc82941377)

[Project Purpose: 2](#_Toc82941378)

[Background Information 2](#_Toc82941379)

[Data Platform: 2](#_Toc82941380)

[Data Processing: 2](#_Toc82941381)

[Model Selection 2](#_Toc82941382)

[Data Preparation and Exploration 3](#_Toc82941383)

[Data Preparation: 3](#_Toc82941384)

[Exploratory Data Analysis: 4](#_Toc82941385)

[Frequency Tables: 5](#_Toc82941386)

[Data Partition (Supervised Model): 8](#_Toc82941387)

[Results 9](#_Toc82941388)

[Model 1 Results: Random Forest 9](#_Toc82941389)

[Model 2 Results: Neural Network 12](#_Toc82941390)

[Conclusions 15](#_Toc82941391)

[Model Performance 15](#_Toc82941392)

[Lessons Learned 15](#_Toc82941393)

[Next Steps 16](#_Toc82941394)

# Introduction

## Project Purpose:

The purpose of this project is to complete a data mining analysis for a career transition program focused on helping veterans and those transitioning out of the military.

## Background Information

The data selected for this analysis is from an organization that is currently assisting transitioning active-duty service members, veterans, and their spouses with career transition. This program is focused industry recognized certifications in IT and project management and assists program participants with finding a job after program completion. While business intelligence reporting has been conducted for various reasons in support of the program, an in-depth data mining analysis to look for unknown insights has never been conducted.

## Data Platform:

The data platform selected for this analysis is SAS Viya. SAS Viya includes various tools that assist with data a mining includes a programing interface which can be used for data processing and machine learning. In addition, it includes a point and click data processing tool called Data Prep and a point and click machine learning tool called Model Studio. For the purposes of this project, the analysis will be conducted with the programming interface for data processing and model creation.

## Data Processing:

The data being used for this project has been processed by the organization and will need very little additional processing. The data includes demographic factors along with course selection and military information. SAS utilizes two data types, character and numeric. While each data types can utilize formats to adjust the look of the data, the data remains the same under neither the format. Continuous variables will be changed into discrete variables, placing them into bins. Model documentation will be reviewed to ensure the data processing matches the models selected for the analysis.

## Model Selection

Based on the available data variables the two models selected for this analysis were:

1. Random Forest
   1. Documentation:
      1. [PROC](https://documentation.sas.com/doc/en/casml/8.3/casml_mbanalysis_toc.htm) [FOREST](https://documentation.sas.com/doc/en/vdmmlcdc/8.1/casml/viyaml_forest_syntax01.htm)
2. Neural Network
   1. Documentation:
      1. [PROC NNET](https://documentation.sas.com/doc/en/vdmmlcdc/8.1/casml/viyaml_nnet_gettingstarted.htm)

# Data Preparation and Exploration

## Data Preparation:

To start the data preparation for this project, a list of variables was selected from a larger dataset available for general use for analytics. 13 attributes (including an ID attribute) were selected for use in the initial model. The dataset contains a combination of course, demographic, and military information about the participants.

|  |  |
| --- | --- |
| **Attribute** | **Definition** |
| id | Identification given to the contact in the database |
| course\_name | Name of the course selected by the participant |
| course\_model | Whether the participant completed the course online or in person |
| rank | Military rank of the participant |
| military\_connection | Type of military connection (how the participant is connected to military service through their own service or through a spouse) |
| military\_branch | Military branch of participant |
| education | Education level of participant |
| gender | Gender of participant |
| age | Age of participant grouped |
| race | Race of participant |
| current\_work\_status | Current work status (whether the participant is employed or not) |
| state | State of residence |
| graduated | Whether a participant graduated or not |

* Military connection is a multi-select question, data was adjusted to select only one option based on their first or only choice
* The data had the word “Missing” for many missing values in the attributes. That word was removed, and a null value put in the place
* Age was grouped by the following:
  + 18-24
  + 25-34
  + 35-44
  + 45-54
  + 55+
* State column was cleaned because this field a text field in the database. A left join was used to join onto the state column and if the state matched a two digit or full state name, a cleaned two-digit state was used. If it did not match, it was considered missing.
* Graduated was given a value of 1 if the participant graduated the course. Otherwise, a zero was given.

Data Contents:

Table

Description automatically generated

Figure : Data Contents

## Exploratory Data Analysis:

The first part of the exploratory data analysis was to determine how many levels are included with each variable. A level is a distinct value for that variable. The missing level is defined as the variable having missing or null values. SAS does not utilize ‘null’ values, and instead uses ‘Missing’ to categorize both missing and null. As seen below, all variables have below 10 levels except course\_name, rank, and state.

Table : Variable Levels and Missing Values

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Levels | Missing Levels | Non-Missing Levels |
| Course\_name | 67 | 0 | 67 |
| Course\_model | 3 | 1 | 2 |
| Rank | 27 | 0 | 27 |
| Military\_connection | 5 | 1 | 4 |
| Military\_branch | 8 | 1 | 7 |
| Education | 6 | 1 | 5 |
| Gender | 4 | 0 | 4 |
| Age | 6 | 1 | 5 |
| Race | 9 | 0 | 9 |
| Current\_work\_status | 4 | 1 | 3 |
| State | 53 | 1 | 52 |
| Graduated | 2 | 0 | 2 |

## Frequency Tables:

Frequency tables were created to explore the numerical spread of each variable.

Table

Description automatically generated Table

Description automatically generatedTable

Description automatically generated

Table

Description automatically generatedGraphical user interface, text, application, table

Description automatically generatedTable

Description automatically generated

Application, table

Description automatically generatedGraphical user interface, application, table

Description automatically generated

Table

Description automatically generated Table

Description automatically generatedTable

Description automatically generated

Table

Description automatically generatedTable

Description automatically generated Table

Description automatically generated

Graphical user interface, application, table

Description automatically generated Table

Description automatically generated

Continued data exploration was conducted by the use of a SAS procedure: [Proc Cardinality](https://documentation.sas.com/doc/en/vdmmlcdc/1.0/casml/viyaml_cardinality_overview01.htm). This procedure completes the following:

* Treats all variables as classification variables and attempts to determine the highest levels of each variable not to exceed a specified limit
* Performs a single pass to determine the limited cardinality of each variable
* Runs with all the input variables in the data table or with a specified list of variables
* Recommends a level (CLASS, INTERVAL, or ID) for each variable. You can override these recommended levels in subsequent steps.

Graphical user interface, application, table, Excel

Description automatically generated

## Data Partition (Supervised Model):

A SAS procedure called [Proc Partition](https://documentation.sas.com/doc/en/vdmmlcdc/8.1/casstat/viyastat_partition_syntax01.htm) was utilized. As suggested by SAS, 70% of the total data was selected by the partition. The number of samples can be seen under the “Number of Samples” column in the table. The partition was completed with the graduated attribute is the predictor variable so a proper samples was selected. An output dataset was completed in which the samples had a flag (variable with a 1 or a 0). The sampled data was given a 1and the non-sampled data had a 0.

Graphical user interface, application, table

Description automatically generated

Figure : Proc Partition Results

# Results

## Model 1 Results: Random Forest

Model Settings:

- Number of trees: 50

- Number of bins: 20

- Minimum leaf size: 5

- Training and validation sets separated by flag variable created earlier

The model results can be seen below. The out of box misclassification rate was over 37%, however that rate dropped below 35% at about 8 decision trees with the training data. Once the decision trees were above 10, there was no visible improvement in the model demonstrating that a simpler model would fit the data much better than a more complex model. Additionally, because the split was so high for training, that seems to have affected the overall success of the model. The training model functioned better than the validation model, however that may increase for the validation data if more observations are included.

Table

Description automatically generated

Figure : Random Forest Results

The below figure shows the improvement in the model misclassification rate with the increase in decision trees. The peak misclassification rate for training was 0.35, and for validation was 0.37.

Chart, histogram

Description automatically generated

Figure : Random Forest Misclassification Rate Graph

The fit statistics can be seen for the both the training and validation data. Like the misclassification rate, the fit statistics have better performance with the training data than the validation data.

Graphical user interface, application, table

Description automatically generated

Figure : Random Forest Training Fit Statistics

Table

Description automatically generated

Figure : Random Forest Validation Fit Statistics

The ROC curve shows the training and validation positive rate to be above the random selection line. While the model has a large amount of error, it did produce better results than random selection.

Chart, line chart

Description automatically generated

Figure : Random Forest ROC Curve

The lift for this model is relatively low with a peak at 1.5.

Chart, line chart

Description automatically generated

Figure : Random Forest Lift Chart

## Model 2 Results: Neural Network

The Neural Network model produced a validation misclassification rate of 0.3706 or just over 37%. The model used approximately 31,000 of the 37,000 observations read due to missing values. The total number of nodes created was 187.

Graphical user interface, application

Description automatically generated

Figure : Neural Network Results

Similar to the random forest, the fit statistics performed better with the trianing data over the validation dataset. However, a much larger amount of observations were use for training vs. validation and may have skewed the results.

Graphical user interface, text, application

Description automatically generated

Figure : Neural Network Training Fit Statistics

Graphical user interface, text, application

Description automatically generated

Figure : Neural Network Validation Fit Statistics

The ROC curve for the neural network model was above the random selection line which is positive.

Chart

Description automatically generated

Figure : Neural Network ROC Curve

The peak lift achieved was 1.4 and steadily declined.

Chart, line chart

Description automatically generated

Figure : Neural Network Lift Chart

# Conclusions

## Model Performance

Overall, both models performed similarly with the data. Both ROC curves as seen below performed better than the random selection for the model, however the Random Forest seems to have performed better. Additionally, the misclassification rate for the Random Forest was higher for the validation data, it dropped to below 0.35 for the training data. The random forest seems to have performed better, especially with the training. The misclassification rate never dropped below the mid-thirties, which is a large amount of error and would need to be improved upon if the model was to be put into production and try to predict the graduates.

Chart

Description automatically generatedChart, line chart

Description automatically generated

Figure : Random Forest ROC Curve (Left) and Neural Network ROC Curve (Right) Comparison

|  |  |  |
| --- | --- | --- |
| Misclassification Rate | Training | Validation |
| Random Forest |  | 0.3754 |
| Neural Network | 0.3616 | 0.3706 |

## Lessons Learned

The models selected for this project were both models in which the model relied upon nominal variables. The models had input for interval, ordinal and continuous, however the data being structured as it was, it did not allow for the use of those variables. Initially a clustering model was selected, however clustering models rely upon continuous, or interval variables and the data did not have any to support use of that model. Additionally, with the models used, the separation of training and validation data allowed for better performance with the training data (70% of the data provided) and performed worse among the validation data (remaining 30%). This demonstrates that the 70/30 split, may not be ideal for the data and a cross-validation method for training the model may have been a better fit for the data.

## Next Steps

* Research the data and determine if any interval variables are available that may provide better insight for clustering.
* Try cross-validation for training rather than the 70/30 method as suggested by SAS. This may cause better performance.
* Utilize other models and test performance to determine which model is the best fit for the data.