Assignment Two

Ben Waetford

2022-08

edX HarvardX PH125.9x

Capstone Project: IDV Learners

Introduction

Purpose

The HardardX Data Science Professional Certificate requires students to complete nine courses. This paper (the Individual Assignment) is intended to satisfy the final requirement of the ninth course. The Individual Assignment requires students to apply machine learning techniques (beyond simple linear regression) to answer a question of their own choice using data that is *not* well known. Thus, for this Individual Assignment, I have chosen to use a data set from a hypothetical record label known as Acme Musical Sounds Inc.

Data

Data Selection The data set employed in this Individual Assignment represents data from Acme Musical Sounds Inc.'s Enterprise Resource System.

Data Definitions Acme Inc. regularly enters into contracts to produce soundtracks for various clients. Each contract is segmented by a number of factors, and the executes projects for a range of clients. The projects are variable across multiple independent variables. One dependent variable, "Result" describes whether Acme Musical Sounds Inc. earned a lot of money (Big win), earned some money (win), broke even (close enough) lost money (loss), or lost a lot of money (big loss) as a result of completing the agreed upon music production.

Variable	Description
Client Base	Each project is associated with one of three different client types.
Contract Type	One of five different contract delivery frameworks employed during contract negotiation.
State	The geographic region that the customer requires the project be delivered within.
Joint Venture	An indicator that displays 'yes' if Acme is the sole owner of the project.
Recurring Revenue	An indicator that displays 'Y' if the project is renewed on a regular basis without the need for the client to go to market and seek competitive bids from other prospective vendors or suppliers.
Delivery Model	One of six different models used to manage the delivery of the project.

Variable	Description
Sector	The group within Acme Musical Sounds Inc. that leads the service delivery on the project.
Service Focus	One of 24 service specialty groups within Acme Musical Sounds Inc. that manages the customer relationship.
Duration	A number that represents the duration of the project. A higher number represents a longer duration.
Year	The year that the project began.
OM	A number from one to 10 that represents the relative size of the job, one being the smallest that Acme Musical Sounds Inc. would perform, 10 being the largest.
Result	One of five outcomes: big loss, loss, close enough, win, big win.
Category	An engineered feature that groups the five 'Result' variables into two groups—'Better than expected' (those jobs with an actual profit that was greater than expected) and 'Worse than expected' (those jobs that lost money).

Research Question

The owner of Acme Musical Sounds Inc. continues to pursue opportunities to grow their business by delivering new projects for new customers. The company wishes to know if out of the box Machine Learning models might successfully predict jobs that will result in a profit margin that is better than expected. A successful prediction, for the purpose of this assignment, is any models with an accuracy score greater than the no information rate (0.720).

Research Activities

Source data was imported, and checked for NAs. The data were then transformed into factors, a dependent feature was engineered. Six classification models were fitted and the accuracy of each was reviewed to determine if any accuracy scores were greater than the no information rate.

Results Summary

Two models performed as well as the ino iformation rate. Three models out-performed the no information rate by a small margin (by 0.01 and 0.02 overall accuracy). One model performed worse than the no information rate. Thus, the null hypothesis, that Acme Musical Sounds Inc.'s data sample has little to no predictive potential, has been accepted.

Future analysis should be performed on a larger data set, should incorporate parameter tuning, apply ensemble methods, and explore options to control prevalence (including balancing the sensitivity and specificity scores). Such additional actions may improve upon the accuracy scores reported in this assignment.

Method and Analysis

Libraries

```
{if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")}
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(readxl)) install.packages("readxl", repos = "http://cran.us.r-project.org")
if(!require(rpart.plot)) install.packages("rpart.plot", repos = "http://cran.us.r-project.org")
if(!require(randomForest)) install.packages("randomForest", repos = "http://cran.us.r-project.org")
if(!require(doParallel)) install.packages("doParallel", repos = "http://cran.us.r-project.org")
if(!require(naivebayes)) install.packages("naivebayes", repos = "http://cran.us.r-project.org")
```

Multi-core Processing

Although the data set used for this Individual Project is small, multicolored processing was enables since the caret 'train' function benefits from improved speed when multicore processing is enabled. The doParallel library was used for this purpose and 20 cores were made available; 20 cores were assigned arbitrarily.

```
registerDoParallel(cl)
```

Data Cleaning and Preparation

Most data were categorical, but coded as continuous (numeric). All data were transformed into factors such that they are understood by R as being categorical data–suitable for analysis by classification trees.

Train and test sets were created. 70% of data were allocated for training and the remaining 30% were used for testing. 70% split was chosen since the data set is small and a larger training set would have resulted in a test set that did not contain all factors/levels present in the training set.

```
  mutate(result = as.factor(Result),
         ext_mgmt = as.factor(External_mgmt),
         client_b = as.factor(Client_base),
         contract_t = as.factor(Client_base),
         contractor_t = as.factor(Contractor_type),
         delivery_m = as.factor(Delivery_model),
         state = as.factor(State),
         sector = as.factor(Sector),
         recur_r = as.factor(Recur_revenue),
         service_f = as.factor(Service_focus),
         syear = as.factor(start_year),
         order_m = as.factor(OM),
         Result = NULL,
         External_mgmt= NULL,
         Client_base = NULL,
         Contract_type = NULL,
         Contractor_type = NULL,
         Delivery_model= NULL,
         State = NULL,
         Sector = NULL,
         Recur_revenue = NULL,
         Service_focus = NULL,
         start_year = NULL,
         OM = NULL)
```

Data Exploration and Insights

The data contains no NAs, and contains 1031 rows.

"'{na.s <- sapply(dat, {function(x) any(is.na(x))})} knitr::kable(na.s)

#How many rows? nrow(dat)

'''{}

Models

Results

Results and Performance

Conclusion

Summary

Impact

Limitations

Future Considerations

Control prevalence

Tune parameters

Increase data sample size