Academic Success dataset exploration

Ruslan Nagimov

2024-06-19

Libraries Including

```
library(readr)
library(ggplot2)
library(tidyverse)
library(dplyr)
library(vcd)
```

Data Exploration

Data Loading

```
## # A tibble: 6 x 38
##      id `Marital status` `Application mode` `Application order` Course
##   <dbl>            <dbl>              <dbl>               <dbl>  <dbl>
## 1     0                1                  1                   1   9238
## 2     1                1                 17                   1   9238
## 3     2                1                 17                   2   9254
## 4     3                1                  1                   3   9500
## 5     4                1                  1                   2   9500
## 6     5                1                 39                   1    171
## # i 33 more variables: `Daytime/evening attendance` <dbl>,
## #   `Previous qualification` <dbl>, `Previous qualification (grade)` <dbl>,
## #   Nacionality <dbl>, `Mother's qualification` <dbl>,
## #   `Father's qualification` <dbl>, `Mother's occupation` <dbl>,
## #   `Father's occupation` <dbl>, `Admission grade` <dbl>, Displaced <dbl>,
## #   `Educational special needs` <dbl>, Debtor <dbl>,
## #   `Tuition fees up to date` <dbl>, Gender <dbl>, ...
```

Unique values in columns

```
##                                    Num of Unique values
## id                                                76518
## Marital status                                        6
## Application mode                                     22
## Application order                                     8
## Course                                               19
```

```
## Daytime/evening attendance                                2
## Previous qualification                                   21
## Previous qualification (grade)                          110
## Nacionality                                              18
## Mother's qualification                                   35
## Father's qualification                                   39
## Mother's occupation                                      40
## Father's occupation                                      56
## Admission grade                                         668
## Displaced                                                 2
## Educational special needs                                 2
## Debtor                                                    2
## Tuition fees up to date                                   2
## Gender                                                    2
## Scholarship holder                                        2
## Age at enrollment                                        46
## International                                             2
## Curricular units 1st sem (credited)                      21
## Curricular units 1st sem (enrolled)                      24
## Curricular units 1st sem (evaluations)                   36
## Curricular units 1st sem (approved)                      23
## Curricular units 1st sem (grade)                       1206
## Curricular units 1st sem (without evaluations)           12
## Curricular units 2nd sem (credited)                      20
## Curricular units 2nd sem (enrolled)                      22
## Curricular units 2nd sem (evaluations)                   31
## Curricular units 2nd sem (approved)                      21
## Curricular units 2nd sem (grade)                       1234
## Curricular units 2nd sem (without evaluations)           11
## Unemployment rate                                        11
## Inflation rate                                           13
## GDP                                                      11
## Target                                                    3
```

Since all features found to have less than 5% of unique values, all of them will be treated as an ordinal or categorical. Thus, Chi-squared test can be applied directly

Chi-squared test

Getting features

Defining funtion

```r
chisqr <- function(df, x) {
  tbl <- table(df[[x]], as.factor(df[["Target"]]))
  xsq <- chisq.test(tbl)
  return(c(xsq$statistic, xsq$parameter, xsq$p.value, assocstats(tbl)$cramer))
}
```

Applying Chi-squared

```
results <- lapply(features, function(x) as.vector(chisqr(df, x)))
```

As a result, vector of Chi-Squared, Degree of Freedom, p-value and Cramer's V produced for each feature ~ target pair

Converting results in DataFrame

```
##                             Chi.squared Degree.of.Freedom     p.value   CramerV
## Marital status                 1728.623                 2  0.0000e+00 0.1062804
## Application mode              13093.957                 2  0.0000e+00 0.2925086
## Application order              2048.948                 2  0.0000e+00 0.1157094
## Course                        17313.816                 2  0.0000e+00 0.3363563
## Daytime/evening attendance     1308.506                 1 7.2691e-285 0.1307694
## Previous qualification         5619.960                 2  0.0000e+00 0.1916328
```

Note: Cramer's Value is taken into account since Chi-Squared values and their corresponding values could be calculated incorrectly.

Correlated features based on Cramer's V test and p-value

```
##                                    Chi.squared Degree.of.Freedom p.value
## Application mode                     13093.957                 2       0
## Course                               17313.816                 2       0
## Previous qualification (grade)       13879.295                 2       0
## Admission grade                      17074.870                 2       0
## Tuition fees up to date              15301.107                 1       0
## Gender                                8342.754                 1       0
## Scholarship holder                   12637.604                 1       0
## Age at enrollment                    14943.756                 2       0
## Curricular units 1st sem (enrolled)  11159.131                 2       0
## Curricular units 1st sem (evaluations) 28741.174               2       0
## Curricular units 1st sem (approved)  58407.076                 2       0
## Curricular units 1st sem (grade)     53889.551                 2       0
## Curricular units 2nd sem (enrolled)  11584.771                 2       0
## Curricular units 2nd sem (evaluations) 30008.489               2       0
## Curricular units 2nd sem (approved)  69581.868                 2       0
## Curricular units 2nd sem (grade)     61491.471                 2       0
##                                      CramerV
## Application mode                   0.2925086
## Course                             0.3363563
## Previous qualification (grade)     0.3011528
## Admission grade                    0.3340273
## Tuition fees up to date            0.4471772
## Gender                             0.3301968
## Scholarship holder                 0.4063971
## Age at enrollment                  0.3124878
## Curricular units 1st sem (enrolled) 0.2700340
## Curricular units 1st sem (evaluations) 0.4333666
## Curricular units 1st sem (approved) 0.6177830
```

```
## Curricular units 1st sem (grade)       0.5934108
## Curricular units 2nd sem (enrolled)    0.2751357
## Curricular units 2nd sem (evaluations) 0.4428180
## Curricular units 2nd sem (approved)    0.6742970
## Curricular units 2nd sem (grade)       0.6338852
```

After applying Cramer's V test and p-value test on values we calculated before, we have a list of features that has a significant effect on Target. This will be used for the Machine Learning Model's convenience.

Saving data

Important Note

Since the number of significance tests conducted and visualizations investigated is extremely low, in addition to primary Machine Learning Model, another model with Recursive Feature Elimination will be composed. Performances will be compared to find the best fitting Model.