Ukrainian Catholic University

Faculty of Applied Sciences

Business Analytics

Teaching For Ukraine Data Analysis

Econometrics course

Author:
Eugene Domeretskyi

1 May 2021

1 Introduction

Teaching For Ukraine is a community that is looking for young talents for the role of teachers in needed Ukrainian villages. The volunteer selection campaign is held twice a year. The request of a company was to analyse the process of applications given a one-period dataset.

In this paper, I am aimed to find the consistent characteristics of applicants, test feature significance and provide useful recommendations. First of all, I focus on finding the columns in data that may be informational duplicates of one another. Second step is to remove them and test significance of the remaining features.

2 Data

This section describes data preparation and assumptions made for the dataset to be more informative.

2.1 Description

Recently, the process of applicant's selection in the TFO company was divided into four steps: posting submissions, interview, interview screening, selection. Therefore, the provided dataset (Excel file) consists of four tables, namely .csv files. The first of them includes mostly text comments. The second, screening, contains the main information about applicants, in the form of 0-3 grade level. The third (interview), includes personal information like expectations, motivation or readiness to move to the village. The last one shows the accepted volunteers and the mark of the test they passed.

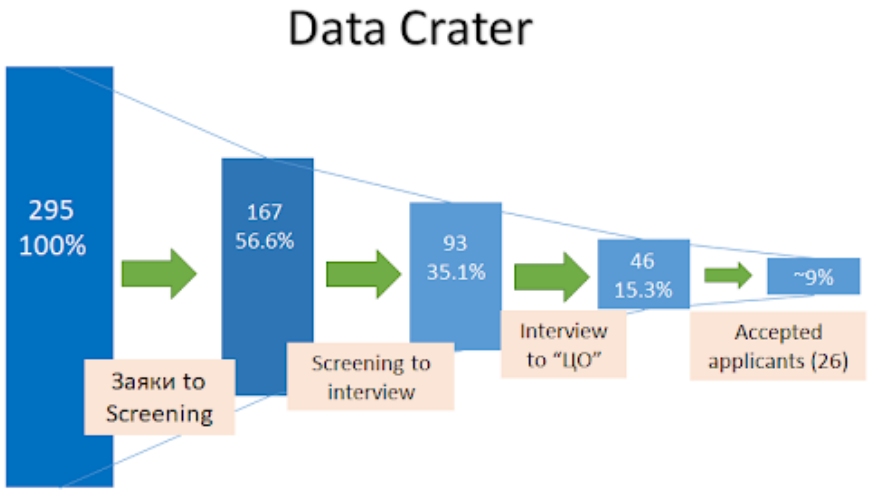

Figure 1: The data flow from one stage to another

Almost all the applicants out of 295 analysed have a bachelor or higher degree (35 master's and 3 doctors). We face the fact that the groups after separation couldn't represent corresponding samples. The natural assumption here is that the applicants should be considered as one group in terms of analysis.

After all, I have the 295 observations to analyse. 167 were accepted to the interview and have summarised data, 93 of them actually came to the interview and 46 was accepted to the “ЦО 1 хвиля”, where only 26 was selected as the future teacher-volunteers.

2.2 Preparation and cleaning

In the project, all the text variables (comments, expectations, personal opinion) were omitted due to the complexity of the analysis.

Several variables like ’Досвід’, ’Мотивація’ have duplicates in different tables, so I left only one of them, the first one, which probably contains the smallest amount of NaN values.

Data merging is performed on the “Name Surname” columns, dropping duplicates and manually checking merge correctness (as data was typed in by humans and even the difference in one character is a fault here). As a result, 295 unique people, with a huge proportion of undefined values in columns because there is no data for those (for example, 295-167=128) applicants who were not interviewed, as well as for the other groups. That begs the question: “How should the missing data be replaced?”.

Here is the assumptions and practical solutions:

The person who wasn’t selected to the interview, has poor understanding of the program and low motivation, so ’Розуміння’, ’Резюме’, "Мотивація", "Розуміння програми" and "Села" should have the min values (from the list of existing entries) instead of NaN.

Undefined "ЗНО" entries will be replaced with 0, as this number represents 0-160 range (200 is max for ЗНО test).

NaN of the "Предметний тест" will be replaced with its min value.

The other features like "Освіта" or "Грамотність" are expected to be normally distributed, and mean value can be assumed in place of missing data.

Also, I created column “Age”, from the DateOfBirth (much better than “Вік” with 1, 2, 3 entries) and transformed “Status” and “Оффер” columns to the binary outcome where 1 indicates “accepted”. Variables “Оцінка” and “Бал” are omitted because they represent the sum of the list of 12 other variables and can threaten by multicollinearity.

3 Methodology explanation

3.1 Logit Model Assumptions

The Logit model is selected for the project bacause we deal with binary outcome, whether a person was accepted or not. For the model to be correct, the following six assumptions should be confirmed:

The response variable is binary: True. In the model, "Оффер" column with "Yes/No" entries was transformed to the binary outcome.

There is a Linear Relationship between explanatory variables and the Logit of the response variable. This assumption is saved and proven with the Box-Tidwell test.

The observations are independent: True. We can assume that all applicants are unrelated to each other as separate individuals.

There is no multicollinearity: this assumption will be proven using the Variance inflation factor in the next two subsections.

No outliers: Considering the fact, that only good applicants allowed to the next stage, wrong individuals are eliminated in each section (for example, when overall grade is less than 26 out of 36 points possible).

The only violation that may arise is that the **sample size is not so big**. Expected probability of the least frequent outcome is assumed to be 1/4 (0, 1, 2, 3 choices). By the rule of thumb, I should have a minimum of 10 cases with the least frequent outcome for each explanatory variable. From here, the maximal suitable amount of features in the model is:

$(10 * n)/0.25 = 295 => n = 295 * 0.25/10 => n$=7

3.2 Correlation

To save the assumption of "No multicollinearity", correlation between variables should be taken into account. The number of observed individuals is not too big and the correlation threshold equal to **0.75** was selected.

Using the correlation plot, we detect and drop three obvious duplicates:

The difference between **'Розуміння'** and 'Розуміння_програми' is that 'Розуміння_програми' is clear independent variable, while 'Розуміння' includes other variables. Therefore, 'Розуміння' will not be useful in the model because it is included in other features.

'Резюме' strongly correlates with 'ЗНО', 'Розуміння', 'Мотивація_х' and 'Розуміння_програми'. This column is a short assumption of those variables and is an example of perfect multicollinearity.

'Команда' correlates with other variables, too and is not important.

The other variables have small correlation and will be used later.

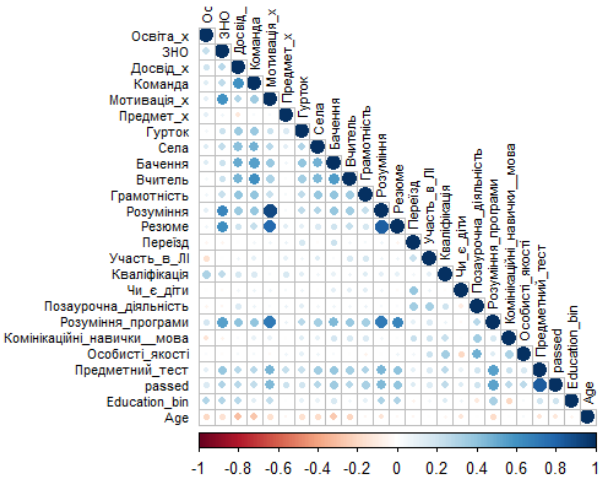

Figure 2: Correlation between all the columns

3.3 Variance inflation factor

For now, data includes 18 variables, and probably, there exists a multicollinearity problem, that may cause p-values unreliability.

The most common way to detect multicollinearity is by using the *variance inflation factor* (VIF), which measures the correlation and strength of correlation between the

predictor variables in a regression model. The idea behind is to regress one variable A on the other B, C, D ... And the higher the R^2 is, the better A is explained by another one, the less we need A.

The VIF value starts from 1 and can go to infinity. As a VIF threshold I used number **10**, as described in *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. 2012 edition.* .

First of all, I removed **"Села"** column because it is explained by "Бачення", "Розуміння програми", and maybe some other variables.

Next, **"Освіта_х"** and "Предметний_тест" are potentail collinear candidates. And I removed "Освіта_х" because it is less informative. The same was done with **"ЗНО"**. These three features are related and answer the same question: *"How can knowledge be measured?"*. The other variables left don't show any significant concerns.

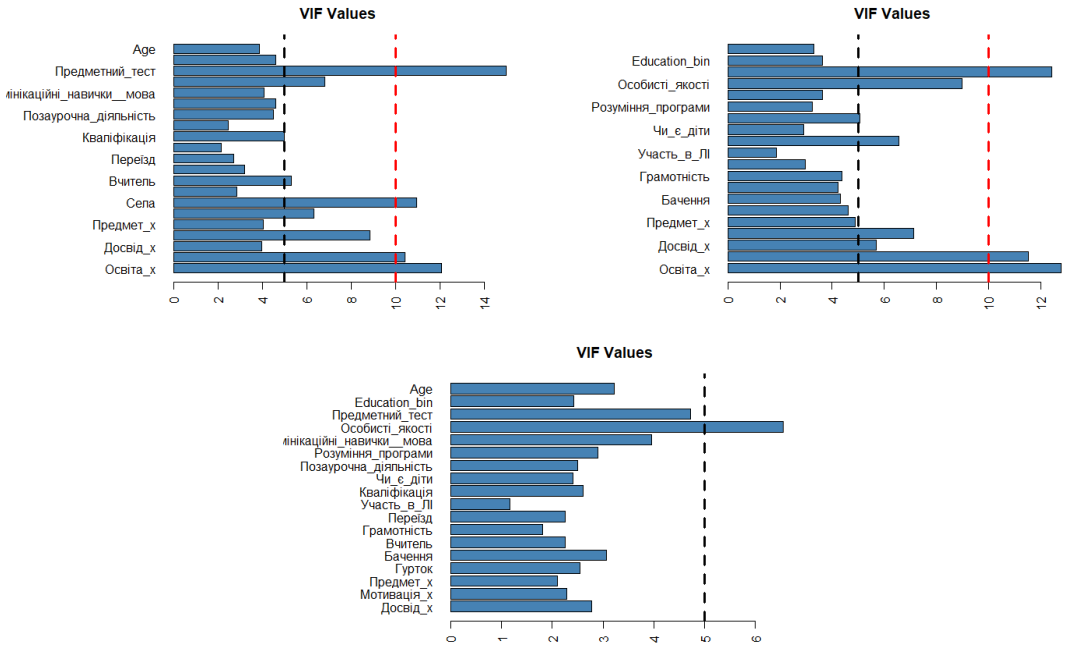

Figure 3: (a) All (b) Освіта_х removed (c) ЗНО removed

3.4 Feature selection

Running the Logit model on the 18 variables we have for now, there are a lot of unimportant features with high p-values. I tried to eliminate them one by one, starting from the most insignificant. In the figure below, this model is called **Straight**. Leaving only "Предметний тест" and "Особисті якості" results in **Small model**.

To avoid testing all possible combinations of them, feature selection was used. This is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

Forward selection: all the variables were left, including 14 insignificant. Model rejected.

Backward selection: There are two insignificant variables with p-val $\eqsim 0.12$, however, the model explained 84% of data with good AIC.

BW improved: insignificant variable "Грамотність" was removed from the original backward model.

Method	AIC	Pseudo R^2	Features with p-val > 0.1	Feature num
Forward	59.412	0.89	10+	18 (all)
Backward (BW)	42	0.84	2	6
BW improved	44.7	0.81	0	5
Straight	45	0.83	0	5

Figure 4: The results of the selection tests.

The summary of these models can be found in the **Appendix** section.

The Akaike information criterion (AIC) is considered the main evaluation criterion for feature selection and the best model should provide the **lowest AIC** value. Long story short, we have two options here to select: backward model model with some insignificant features, or the same, but with significant only.

4 Results

Model with features obtained using backwards feature selection with "Грамотність" removed is the optimal solution to that problem. It meets all the required assumptions (multicollinearity absence, significant variables, low AIC and good R^2).

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -53.39358   11.71421  -4.558 5.16e-06 ***
Предмет_х                     2.26398    1.11776   2.025  0.04282 *
Чи_є_діти                     3.10729    1.15335   2.694  0.00706 **
Комінікаційні_навички__мова   3.78762    1.74036   2.176  0.02953 *
Особисті_якості               2.96602    1.20943   2.452  0.01419 *
Предметний_тест               0.27323    0.05538   4.934 8.06e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.940  on 294  degrees of freedom
Residual deviance:  32.689  on 289  degrees of freedom
AIC: 44.689

Number of Fisher Scoring iterations: 9

[1] "Pseudo R^2"
[1] 0.814206
```

The model includes five independent features:

Предмет: binary variable where two outcomes are possible: a person can teach technical topic (biology, chemistry, physics, math, IT, English) or any others. From here, we can state that the TFO company needs teachers from the first category, so the increase in the log odds is explained.

Чи є діти: 3 indicates no children, 0 indicates children + divorced. If the person has no children, it is much easier for him to move in the new place.

Комунікаційні навчики, Особисті якості are important factors and are preferred, too.

Предметний тест is the most significant feature as it correctly shows the knowledge of a person without any bias.

5 Conclusions

In the project, I made a data review, provided recommendations on how to improve it by replacing missing values, figured out insignificant features and prepared a set of truly reliable ones.

The final model results are quite good and convincing. Starting from 24+ different features, only 5 or 6 of them are really important. That is 'Предмет', 'Діти', 'Комунікаційні навички', 'Особисті якості', 'Предметний тест' and, optionally, 'Грамотність'. With them, even quite good predictions can be made (22 out of 26 were correctly predicted).

The further steps is to collect more data, test it again, maybe on groups divided by scientific degree. The challenge is, naturally, to find the better ways to replace the missing data values and make analysis year-by-year to see whether major world-factors, like pandemic, influence the analysis results. Quick hypotheses, like "Comment length value is non-zero" were rejected in the project. We can try better and make semantic analysis.

6 Appendix

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -63.91479   22.70682  -2.815 0.004881 **
Досвід_х                   0.77334    1.38369   0.559 0.576235
Мотивація_х               -0.04464    1.10848  -0.040 0.967876
Предмет_х                  2.00156    1.66354   1.203 0.228902
Гурток                    -1.73206    1.30156  -1.331 0.183269
Бачення                    0.29388    1.62371   0.181 0.856374
Вчитель                    0.24453    1.11256   0.220 0.826036
Грамотність                1.83759    1.18154   1.555 0.119887
Переїзд                    2.47993    2.10360   1.179 0.238437
Участь_в_ЛІ               -0.03771    0.12964  -0.291 0.771105
Кваліфікація              -0.05684    1.72225  -0.033 0.973674
Чи_є_діти                  3.17582    1.88831   1.682 0.092601 .
Позаурочна_діяльність     -1.42012    2.39547  -0.593 0.553292
Розуміння_програми        -0.38448    0.99528  -0.386 0.699271
Комінікаційні_навички__мова  2.97940    3.34983   0.889 0.373779
Особисті_якості            5.83574    2.64770   2.204 0.027519 *
Предметний_тест            0.31362    0.09497   3.302 0.000959 ***
Education_bin             -1.63294    2.54844  -0.641 0.521679
Age                       -0.05393    0.19697  -0.274 0.784233
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.940  on 294  degrees of freedom
Residual deviance:  23.554  on 276  degrees of freedom
AIC: 61.554

Number of Fisher Scoring iterations: 10

[1] 0.8661244
```

Figure 5: Forward feature selection

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -55.85768   13.64732  -4.093 4.26e-05 ***
Предмет_х                  1.79390    1.19271   1.504   0.1326
Грамотність                1.61281    1.00833   1.599   0.1097
Чи_є_діти                  2.80143    1.35527   2.067   0.0387 *
Комінікаційні_навички__мова  3.83901    1.90945   2.011   0.0444 *
Особисті_якості            3.29370    1.30661   2.521   0.0117 *
Предметний_тест            0.26387    0.06527   4.042 5.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.940  on 294  degrees of freedom
Residual deviance:  28.019  on 288  degrees of freedom
AIC: 42.019

Number of Fisher Scoring iterations: 9

[1] 0.8407451
```

Figure 6: Backward feature selection

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -40.43154    8.47932  -4.768 1.86e-06 ***
Комінікаційні_навички__мова   2.66164    1.47608   1.803   0.0714 .
Грамотність                2.33968    1.12111   2.087   0.0369 *
Особисті_якості            2.88397    1.10244   2.616   0.0089 **
Предметний_тест            0.25154    0.04784   5.258 1.46e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.940  on 294  degrees of freedom
Residual deviance:  35.035  on 290  degrees of freedom
AIC: 45.035
```

Figure 7: Straight elimination feature selection

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -29.65115    4.80642  -6.169 6.87e-10 ***
Особисті_якості   3.29762    1.02419   3.220  0.00128 **
Предметний_тест   0.27300    0.04197   6.504 7.80e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.94  on 294  degrees of freedom
Residual deviance:  46.21  on 292  degrees of freedom
AIC: 52.21

Number of Fisher Scoring iterations: 7

[1] 0.7373549
```

Figure 8: Only significant features left

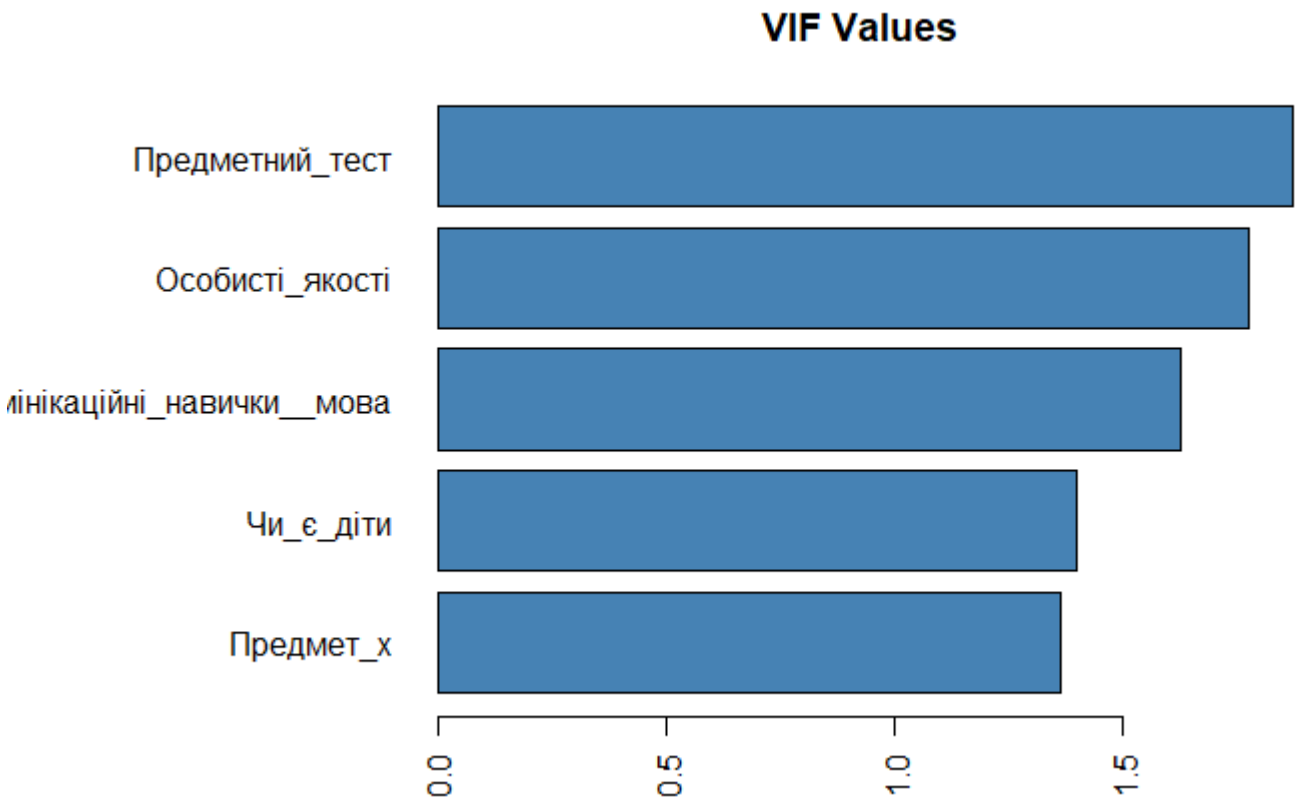

Figure 9: Final model doesn't show covariance concerns