HR CASE

Summer School in Research Methods

Giulechitsa

2019

# HR Case

## Plan of Attack

1. Business understanding

a. Variable review

b. Variable filter

2. Data exploration

a. What do the variables show

b. Descriptives (Stats)

c. Implicit data gathering & fill, data gathered to assumptions

* Бизнес логика / анализ - HR martket specific
* since it is a call center job, we more heavily focused on the language factor
* assumptions - at least one person per add
* focus on the given candidates qualities (not external factors as employer, position of the interviewee)
* we accept that we will model with the people sent to interviews. We have too few cases of people not sent and recorded to verify the HR bias.
* Nationality imputing from score (lang\_grade) -> 7 for native
* is national -> score (lang\_grade) imputing from nationality if native to the position
* Requested salary does not impact sent/filter criteria
* Model that shows which candidates have a better propensity to be sold
* Data prep -> Tree (most likely)
* Да тестваме по работодател по минимум 50 наблюдения -> 10 в листо -> 5-6 над 180 -> може да тестваме модел
* Обучителната и тестовата извадка (train/test split) са определени в данните

3. Data prep

a. Imputing & filters - we used clusterization to recieve the mean values (for each cluster) and then imputed the data with randomized mean values. We took into account the most important variables uncluding age, gender etc.

b. Features - we used correlations & distributions

4. Feature engineering - created a list of new variables - success rate

a. Brainstomring

b. Feature testing

5. Model construction

a. Tree - done in R and Orange

b. Logistic regression - done in Orange

c. Naive - done in Orange

6. Testing - test results pending

a. Feature testing

b. Model testing

c. Sensitivity analysis

d. Model evaluation

7. Implementation

Среда за файлове и документация -> Google Drive

## DataPrep

**Входни данни**: 67 променливи

**Допълнени и преработени**:

* 9 преработени
  + допълване на данни на базата на повтарящи се записи за един клиент
  + допълване на данни на базата на усреднени и зашумени стойности по клъстери
  + бинаризиране
  + интервални скали
* 16 нови променливи
  + Преизчислени стойности на база на други съществуващи променливи (за намаляване на обема на променливите)
  + Коефициенти между бизнес-значими променливи

**Премахнати**: 49 променливи

* 34 поради преизчисление на стойностите им в нови променливи (оценки по езици)
* 15 по други причини (бизнес-логика, ниска дисперсия, индекси)

**Подадени към модела** - **☑** : 34 променливи

* от тях използвани (feature selection)- **☑ ☑:** 24 променливи

|  |  |  |  |
| --- | --- | --- | --- |
| Наименование | Описание | Преработка | Използва се в модела |
| [rec\_id] | уникален индекс на наблюденията |  | 🗷 |
| [candidate\_id] | ID на кандидата |  |  |
| [campaign\_id] | ID на кампанията |  |  |
| [database] | от коя база данни е кандидатът (local/external) |  |   Недостатъчна дисперсия |
| [sex] >  [sex(numeric)] | пол (m/f) | Трансформирана на [sex(numeric)]  F - 0  m - 1 | **☑ ☑** |
| [age\_calc] >  [Age new (cleaned up)] | възраст към момента на интервюто | Преработена в: [Age new (cleaned up)]  Допълнително запълване на данни:  1. От други налични записи на същото [candidate\_id]  2. С усреднени зашумени стойности по клъстери (salary/position/experience) | **☑** |
| *НОВО:* [age\_squared] |  | Изчислената възраст на квадрат | **☑ ☑** |
| *НОВО:* [Native check (native\_1/non\_native\_0)] |  | Натурален носител на езика, за който кандидатства (изчислено на база на оценка 7 за езика, на който кандидатства)  native\_1/non\_native\_0 | **☑** |
| [Nationality] >  [Nationality / Native] | националност на кандидата (German, Italian, Bulgarian, etc.) | Допълнена иформация за натурален носител на езика (оценка по езика – 7) - [Nationality / Native] | **☑ ☑** |
| *НОВО:* [Nationality / Native\_bin] | националност на кандидата (German, Italian, Bulgarian, etc.) – в числова стойност | Допълнтелна преработка на [Nationality / Native]  Чрез разделяне в 4 категории на базата на средната успеваемост (result) на категорията:  'Native Bulgarian Position native speaker' - 0.4000  'Foreigner Position native speaker' - 0.2105  'Foreigner Non-native to the position' - 0.1460  'Native Bulgarian Non-native to the position' - 0.1715 | **** |
| [experience] > [experience\_new\_cleanup] | опит на кандидата в месеци | [experience\_new\_cleanup]  Допълнително запълване на данни:  1. От други налични записи на същото [candidate\_id]  2. С усреднени зашумени стойности по клъстери  (salary/position/age) | **☑ ☑** |
| [salary] > [salary\_new\_cleanup] | очаквана заплата | Преработена в [salary\_new\_cleanup]  Допълнително запълване на данни:  1. От други налични записи на същото [candidate\_id]  2. С усреднени зашумени стойности по клъстери  (age/position/experience) | **☑ ☑** |
| *НОВО*: [salary/native] |  | Коефициент между очакваната основна заплата и натурален носител на езика („колко струва човекът в зависимост от това дали е натурален носител на езика) | **☑** |
| *НОВО*: [salary per 1 lg grade] |  | Коефициент между очакваната основна заплата и оценката по първия език („колко струва човекът в зависимост от знанията му по езика“) | **☑ ☑** |
| [available\_to\_start] > [available\_to\_start\_bin] | брой седмици след които кандидатът може да започне работа | Оригинално – в седмици  Променена на [available\_to\_start\_bin]:  - 0, 1 седмици - 1  - 2, 3, 4 седмици - 2  - повече от 5 седмици - 3 | **☑ ☑** |
| [source\_name] > source\_name by group] | източник на кандидата (direct, referral, etc.) | Променена на [source\_name by group]  Оригинално – в 6 групи  Преработено в 2 групи – Direct/others | **☑ ☑** |
| [grade\_name] > [senior\_dummy\_logit] | йерархия на позицията (regular, junior, senior, etc.) | Променена на [senior\_dummy\_logit]  Оригинално – в 5 групи  Преработено в 2 групи - Seniors/Others | **☑** |
| [wtime\_name] | предпочитания за работно време |  | **☑ ☑** |
| [position\_type] | 3 типа позиция (сold call, support, analyst) |  |   Недостатъчна дисперсия |
| [employer\_name] | име на компанията работодател |  | **☑**  Допълнителни модели |
| [technical\_position] | наличие на изискване за технически познания (yes, no) |  |   Недостатъчна дисперсия |
| [lead] | отговорникът за конкурса, който контактува с клиента |  |  |
| [contactp\_name] | контактното лице за кампанията |  |  |
| [weekday\_presented] | ден от седмицата, в който е презентиран |  |  |
| [monthday\_presented] | ден от месеца, в който е презентиран кандидатът |  |  |
| [month\_presented] | месец, в който е презентиран кандидатът |  |  |
| [quarter\_presented] | тримесечие, в което е презнтиран кандидатът |  |  |
| [year\_presented] | годината, в която е презентиран кандидатът |  |  |
| [interviews\_count] | брой интервюта, на които се е явявал кандидата |  | **☑ ☑** |
| *НОВО*: [Num\_Successes] |  | Брой успешни интервюта на кандидата | **☑** |
| *НОВО*: [Num\_Failures] |  | Брой неуспешни интервюта на кандидата | **☑ ☑** |
| [employees\_needed] | брой търсени служители по настоящата кампания |  |  |
| [filter] | обобщаващ филтър за статуса на кандидата: (no, in process, not sent, yes) |  | **☑**  Включени са само кандидати с филтър NO |
| [result] | target variable (0/1)  резултат от селекцията (назначен) |  | **☑**  target variable |
| [sample] | Train/Test |  | **☑** |
| [lang\_name] | език, необходим за позицията |  | **☑ ☑** |
| [lang\_1] >  [lang\_1\_bin] | първи език на кандидата | Оригинално: 23 езика  Преработено в 6 групи на базата на брой кандидати по език на [lang\_1\_bin]:  'German'  'French'  'English'  'Spanish'  'Italian'  'Other' | **☑ ☑** |
| *НОВО*: [lang\_1\_grade] |  | Оценка по първия език – преизчислено от индивидуалните оценки по езиците | **☑ ☑** |
| [lang\_2] | втори език на кандидата |  | **☑ ☑** |
| *НОВО*: [lang\_2\_grade] |  | Оценка по втория език – преизчислено от индивидуалните оценки по езиците | **☑ ☑** |
| [lang\_3] | трети език на кандидата |  | **☑** |
| *НОВО*: [lang\_3\_grade] |  | Оценка по третия език – преизчислено от индивидуалните оценки по езиците | **☑ ☑** |
| [lang\_grade] | оценка на езика за който кандидатства |  | **☑ ☑** |
| *НОВО*: [Languages zusammen] |  | Текстов стринг, включващ комбинация от всички езици, които владее кандидата (English Italian Spanish) | **☑ ☑** |
| *НОВО*: [Number of languages spoken] |  | Брой езици, които владее кандидата | **☑ ☑** |
| *НОВО*: [Number of languages spoken\_bin] |  | Брой езици, които владее кандидата -  Допълнително преработен в  [Number of languages spoken\_bin]  1 език - 0  Повече от 1 език - 1 | **☑ ☑** |
| *НОВО*: [Average grade of spoken languages] |  | Средна оценка от всички езици, посочени в кандидатурата | **☑ ☑** |
| *НОВО*: [Average score of other languages (not lang\_name) |  | Средна оценка от останалите езици, които владее | **☑ ☑** |
| *НОВО*: [result\_lag] |  | Дали на предишното кандидатстване кандидатът е бил приет или не  1 – да  0 – не  Null – няма предишни кандидатствания | **☑ ☑** |
| [Arabic] | оценка арабски | Всички индивидуални оценки по езици са преработени като оценки по първи, втори и трети език |  |
| [Armenian] | оценка арменски |  |  |
| [Bulgarian] | оценка български |  |  |
| [Chinese] | оценка китайски |  |  |
| [Croatian] | оценка хърватски |  |  |
| [Czech] | оценка чешки |  |  |
| [Danish] | оценка датски |  |  |
| [Dutch] | оцека холандски |  |  |
| [English] | оценка английски |  |  |
| [Estonian] | оценка естонска |  |  |
| [Finnish] | оценка финландски |  |  |
| [French] | оценка френски |  |  |
| [German] | оценка немски |  |  |
| [Greek] | оценка гръцки |  |  |
| [Hebrew] | оценка иврит |  |  |
| [Hungarian] | оценка унгарски |  |  |
| [Italian] | оценка италиански |  |  |
| [Japanese] | оценка японски |  |  |
| [Kazakh] | оценка казакски |  |  |
| [Korean] | оценка корейски |  |  |
| [Latvian] | оценка латвийски |  |  |
| [Lithuanian] | оценка литовски |  |  |
| [Nordics] | оценка скандинавски езици |  |  |
| [Norwegian] | оценка норвежки |  |  |
| [Polish] | оценка полски |  |  |
| [Portuguese] | оценка португалска |  |  |
| [Romanian] | оценка румънски |  |  |
| [Russian] | оценка руски |  |  |
| [Serbian] | оценка сръбски |  |  |
| [Slovak] | оценка словашки |  |  |
| [Slovenian] | оценка словенски |  |  |
| [Spanish] | оценка испанска |  |  |
| [Swedish] | оценка шведски |  |  |
| [Turkish] | оценка турски |  |  |