**HR CASE**

**Summer School in Research Methods**

**Giulechitsa**

**2019**

**Съдържание:**

[1. Въведение 3](#_Toc12216765)

[2. Бизнес проблем 3](#_Toc12216766)

[3. Разглеждане на данните 3](#_Toc12216767)

[4. Обработка на данните 5](#_Toc12216768)

[5. Моделиране и резултати 13](#_Toc12216769)

# Въведение

Целта на текущия документ е да опише методологията и стъпките за решаване на задачата „HR Case“. Базиран е на CRISP-DM структурата, следваща фазите:

1. Разбиране на бизнес проблема и правилно дефиниране на задачата;
2. Разбиране на наличните данни;
3. Обработка и обогатяване на данните;
4. Създаване и тестване на модели за решение на задачата;
5. Избиране на най-добър модел и тълкуване на резултатите.

# Бизнес проблем

В проекта се разглежда основната дейност на анонимна компания за човешки ресурси (***XXX***), специализирана в обслужване на кол центрове. Във всекидневната си работа служителите ѝ оценяват всички явили се при тях кандидати, като се опитват да ги подберат така, че да бъдат посрещнати нуждите и желанията на компаниите-клиенти. Ефективността на процеса на подбор се оценява чрез резултите на изпратените на интервю кандидати – дали им е предложена оферта, или не. Точните критерии на подбора не са изяснени, но е изрично упоменат фактът, че към фирмите-работодатели се изпращат единствено претендентите с оценка поне 5 на изпита по изискания в обявата език. Това предполага съществуването на имплицитно изместване на извадката, тъй като решението за изпращаните кандидати се базира на недостъпна за анализаторите информация.

Първата ясно дефинирана задача е да се конструира модел, който да различава с възможно най-голяма точност перспективните кандидати. За ***ХХХ*** загубата на потенциално избран претендент е много по-скъпа от изпращането на претендент, който в последствие е отхвърлен. Затова при избирането на модел е важно да се постави акцент върху намирането на истинските единици (т.е. назначените лица), а не толкова върху намаляването на сгрешените единици и увеличение на правилните нули. Принципът е илюстриран в ***Таблица 1***.

Таблица 1: Матрица на резултатите

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Прогнозиран резултат** | |
|  |  | **0** | **1** |
| **Реален резултат** | **0** | Правилна оценка | Допустима грешка |
| **1** | Скъпа грешка | Правилна оценка |

Втора основна задача е да се провери хипотезата дали различните компании имат коренно различни стратегии при избиране на кандидати. Според ***ХХХ*** е възможно в някои от компаниите-клиенти да има субективни фактори, които водят до нисък процент в реализацията на претендентите. За целта е необходимо да се изработят индивидуални обясняващи модели за кол центровете, за които е възможно. Разликата в значимите показатели би означавало, че в ***ХХХ*** би могло да се изработят стратегии за по-добър подбор на лица.

# Разглеждане на данните

Процеса по разбиране на данните включва множество стъпки. Следват бележки от първата среща на работната група, описващи плана за разплитане на основните 2 цели на проекта.

*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*

Заради многостранното развитие на членовете в екипа бе взето решението задачата да се разработи на няколко софтуера и платформи: Excel, SPSS Modeler, Orange, Jupyter (Python), RStudio.

Разглеждайки данните, бяха забелязани следните характерности:

1. Данните са на ниво кандидатура.
2. Едно лице може да е кандидатствало повече от веднъж.
3. Всяка кампания за търсене на персонал може да има повече от една позиция.
4. Една позиция се дефинира като уникална комбинация между полетата ***campaign\_id*** (ID на кампанията), ***grade\_name*** (йерархия на позицията), ***position\_type*** (тип на позицията), ***employer\_name*** (име на компанията работодател), ***lang\_name*** (език, необходим за позицията), ***lead*** (отговорникът за конкурса, който контактува с клиента), ***technical\_position*** (наличие на изискване за технически познания).
5. Някои от регистрираните кампании продължават години. Това означава, че данните не са изчистени и е възможно множество независими реални кампании за търсене на персонал да са обединени с общо ID.
6. От близо 2000-те наблюдения в базата, само около 1300 могат да бъдат използвани за моделиране, тъй като само те отговарят на условието да са ***реално изпратени на фирмите кандидатури, за които има реален резултат***. За целта се използва полето ***filter = “no”***.
7. В повечето колони, съдържащи информация за оценките по езици, необходими езици за позициите, произход и години на кандидатите могат да бъдат допълнени редове през засичане на ID на кандидатите, компаниите и кампаниите.

Описаните открития доведоха до следните изводи:

1. За моделиране ще бъдат използвани класически и прости методи, които не изискват наличие на големи данни. Към моделите от този тип, подходящи за прогнозиране на вероятности (за бинарни целеви променливи) са: logistic regression, Naïve Bayes, decision tree, random forest.
2. Анализът за Задача 1 ще бъде изготвен на ниво кандидатура. По този начин ще се улови всяко състояние на потенциални кандидати. Ще се разглеждат променливи, описващи самите кандидати. Идеята е накрая моделът да може да направи генерални изводи за цялата клиентска база на ***ХХХ*** – настояща и бъдеща. Същевременно няма да се вземат предвид променливи, описващи компаниите, кампаниите и позициите, към които са кандидатствали лицата.
3. Анализът на Задача 2 ще бъде на ниво кандидатура, но самите модели ще бъдат индивидуални за отделните компании, за които има достатъчно наблюдения. В анализа ще се вземат предвид и характеристиките на компаниите, кампаниите и позициите.

# Обработка на данните

В началото на фазата входните данни съдържат 67 променливи. От тях допълнени и преработени са 9. Обработката включва следните операции:

* допълване на данни на базата на повтарящи се записи за един клиент;
* допълване на данни на базата на усреднени и зашумени стойности по клъстери. Клъстерите са генерирани на база полета, свързани с допълваните колони;
* групиране на числовите променливи в бинове чрез анализ на разпределението на въпросните променливи;
* закодиране на категориите на някои от колоните със средния резултат за целевата променлива в съответната категория.

Създадени са 16 нови променливи. Два от основните метода са преизчисляване стойностите на база на други съществуващи променливи (за намаляване на обема на променливите) и създаване на коефициенти между бизнес-значими променливи (през отношения, тежести, мултиплициране). Освен това за премахнати 49 променливи:

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

В резултат към модела са подадени 34 променливи, като след анализ на корелационната матрица, са ограничени на 24. Матрицата с корелации се намира в прикачения документ:



***Таблица 2*** описва подробно данните, както и направените преобразувания. Използвани са следните означения:

* Подадени към модела - ☑(34 променливи);
* Използвани (след feature selection)- **☑ ☑ (**24 променливи).

Таблица 2: Преработка и създаване на променливи

|  |  |  |  |
| --- | --- | --- | --- |
| **Колона** | **Описание** | **Преработка** | **Използва се в модела** |
| [rec\_id] | уникален индекс на наблюденията |  | 🗷 |
| [candidate\_id] | ID на кандидата |  | ý |
| [campaign\_id] | ID на кампанията |  | ý |
| [database] | от коя база данни е кандидатът (local/external) |  | ý  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 | **ý** |
| [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) |  | ý  Недостатъчна дисперсия |
| [employer\_name] | име на компанията работодател |  | **☑**  Допълнителни модели |
| [technical\_position] | наличие на изискване за технически познания (yes, no) |  | ý  Недостатъчна дисперсия |
| [lead] | отговорникът за конкурса, който контактува с клиента |  | ý |
| [contactp\_name] | контактното лице за кампанията |  | ý |
| [weekday\_presented] | ден от седмицата, в който е презентиран |  | ý |
| [monthday\_presented] | ден от месеца, в който е презентиран кандидатът |  | ý |
| [month\_presented] | месец, в който е презентиран кандидатът |  | ý |
| [quarter\_presented] | тримесечие, в което е презнтиран кандидатът |  | ý |
| [year\_presented] | годината, в която е презентиран кандидатът |  | ý |
| [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 |  | ý |
| [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 | ý |
| [Armenian] | оценка арменски |  | ý |
| [Bulgarian] | оценка български |  | ý |
| [Chinese] | оценка китайски |  | ý |
| [Croatian] | оценка хърватски |  | ý |
| [Czech] | оценка чешки |  | ý |
| [Danish] | оценка датски |  | ý |
| [Dutch] | оцека холандски |  | ý |
| [English] | оценка английски |  | ý |
| [Estonian] | оценка естонска |  | ý |
| [Finnish] | оценка финландски |  | ý |
| [French] | оценка френски |  | ý |
| [German] | оценка немски |  | ý |
| [Greek] | оценка гръцки |  | ý |
| [Hebrew] | оценка иврит |  | ý |
| [Hungarian] | оценка унгарски |  | ý |
| [Italian] | оценка италиански |  | ý |
| [Japanese] | оценка японски |  | ý |
| [Kazakh] | оценка казакски |  | ý |
| [Korean] | оценка корейски |  | ý |
| [Latvian] | оценка латвийски |  | ý |
| [Lithuanian] | оценка литовски |  | ý |
| [Nordics] | оценка скандинавски езици |  | ý |
| [Norwegian] | оценка норвежки |  | ý |
| [Polish] | оценка полски |  | ý |
| [Portuguese] | оценка португалска |  | ý |
| [Romanian] | оценка румънски |  | ý |
| [Russian] | оценка руски |  | ý |
| [Serbian] | оценка сръбски |  | ý |
| [Slovak] | оценка словашки |  | ý |
| [Slovenian] | оценка словенски |  | ý |
| [Spanish] | оценка испанска |  | ý |
| [Swedish] | оценка шведски |  | ý |
| [Turkish] | оценка турски |  | ý |

# Моделиране и резултати

За **Задача 1** са използвани следните модели в Orange:

1. Tree – с ограничение дървото да не е бинарно. Поставени са ограниченията:
   1. Поне 30 наблюдения в последните листа;
   2. Да не се разделят групи под 60 наблюдения;
   3. Да не се генерират повече от 100 нива на разделение.
2. Random forest с:
   1. До 10 дръвчета;
   2. Ограничение да не се разделят групи под 5 наблюдения
3. Naïve Bayes
4. Logistic regression – Ridge (L2) с параметър C = 1.

Резултатите от моделите са описани в ***Таблицa 3***:

Таблица 3: Представяне на тестваните модели

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Резултати върху данни за обучение | | | | | |
| **Method** | **AUC** | **CA** | **F1** | **Precision** | **Recall** |
| Random Forest | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Tree | 0.91 | 0.82 | 0.83 | 0.80 | 0.85 |
| Logistic Regression | 0.80 | 0.72 | 0.72 | 0.71 | 0.73 |
| Naive Bayes | 0.75 | 0.67 | 0.65 | 0.69 | 0.62 |
|  |  |  |  |  |  |
| Резултати върху данни за тестване | | | | | |
| **Method** | **AUC** | **CA** | **F1** | **Precision** | **Recall** |
| Random Forest | 0.64 | 0.81 | 0.18 | 0.30 | 0.13 |
| Logistic Regression | 0.63 | 0.62 | 0.32 | 0.23 | 0.54 |
| Naive Bayes | 0.58 | 0.59 | 0.26 | 0.19 | 0.42 |
| Tree | 0.53 | 0.63 | 0.26 | 0.19 | 0.38 |

Макар random trees моделът да има най-добро представяне по всички показатели, в данните за обучение се вижда, че се напасва прекалено много към тях (overfitting). Същевременно матрицата с познати и сгрешени резултати показва, че консистентно логистичната регресия познава повече реално наети кандидати (***Таблица 4***). Както е описано по-горе, това е групата от по-голям интерес за компания ***ХХХ***, затова и за модел-шампион се избира логистичната регресия.

Таблица 4: Матрица с попадения и грешки



Заради липса на времеви ресурси, работната група не успя да завърши **Задача 2**. Все пак при бърза проверка през логистични регресии към трите компании с най-много наблюдения се установиха следните зависимости:

1. ***Comp12*** предпочита кандидати с повече опит и познания, повече езици и местни за необходимия език за позицията.
2. ***Comp4*** избира повече млади кандидати с по-малко опит и познания по езици.
3. ***Comp22*** предпочита кандидати с повече езици и по-ниска изисквана заплата. Освен това има значение продължителността на самата кампания, както и отговорното лице по кампанията. Може би става въпрос за намеса на субективни фактори при назначаване на кандидатите.