

HR аналитика

Коста Грујчић

12/2017

Сажетак

Циљ пројекта је истренирати логистичку регресију над подацима о предвиђању одласка запосленог из предузећа на основу великог броја параметара за сваког од њих.

1 Увод

Како је у питањеу проблем бинарне класификације, природно је користити метрику прецизности за модел, док се успешност над појединачним класама мери односом одзива и прецизности.

Приметимо да нам је од веће важности предвидети да ће неки запослени напустити предузеће него да он неће. Уколико за запосленог тврдимо да ће он напустити предузеће, његов надређени или неко из HR тима може обавити додатни разговор са њим и тиме утврдити евентуални проблем. С друге стране, уколико за некога тврдимо да ће остати, он одлуку о одласку доноси изненада. С тим у вези, грешке прве и друге врсте нису еквивалентне. Зато је пожељно максимизирати одзив предвиђања одласка запосленог. Како нам је ипак важно да и прецизност буде што је могуће већа, коначан циљ је максимизирати F_1 меру те класе.

Надаље ћемо одлазак запосленог звати позитивном класом, јер нам је циљ предвидети је, док ћемо останак запосленог звати негативном класом.

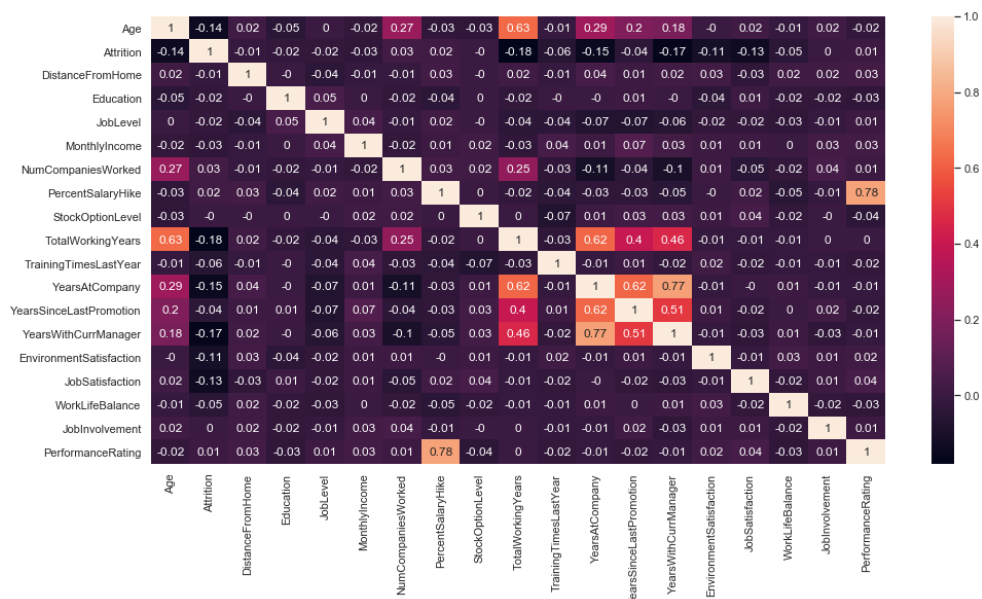
2 Подаци

Подаци који су доступни су подељени у три документа, али како је сваки запослени представљен јединственим идентификатором лако их

можемо спојити у један скуп података. Тако добијени скуп података третирамо као табелу чије су колоне називи атрибута, док редове тумачимо као вишедимензионе векторе. Увидом у податке можемо закључити следеће:

- Колоне OVER18, STANDARDHOURS и EMPLOYEECOUNT можемо уклонити јер су једнаке за све редове у табели.
- Присутни су недостајући подаци у појединим колонама.
- Постоје нумерички и категоријски атрибути.
- Негативна и позитивна класа нису равномерно заступљене.

Из матрице корелације (слика 1) можемо видети да велики број атрибута нису у корелацији, док и када она евидентно постоји није значајно велика. То додатно отежава одабир релевантних атрибута јер међу њима постоји нелинеарна зависност.



Слика 1: матрица корелације нумеричких атрибута

3 Приступ

Како је у питању класификациони проблем са две класе, применићемо логистичку регресију. Подаци који су нам на располагању садрже

велики број атрибута од којих су неки очекивано ирелевантни. Будући да логистичка регресија не може бити директно примењена на категоријске атрибуте, потребно је извршити њихову трансформацију. Зато ћемо описати одабир својстава и начин трансформисања категоријских атрибута.

Оба поменути проблема се могу решавати независно, али и заједно, те ћемо описати оба поступка и упоредити их. Надање ћемо их редом звати *експлицитни* и *имплицитни* приступ.

3.1 Увођење помоћних атрибута и експлицитно одстрањивање ирелевантних атрибута

У овом приступу је категоријске атрибуте неопходно пресликати у нумеричке. У ту сврху се такви атрибут пресликавају у ретке векторе који на само једном месту имају јединицу, док су свуда остало нуле. Димензија тог вектора је једнака броју вредности које одговарајући категоријски атрибут може имати.

Конкретно, ако категоријски атрибут има свега две могуће вредности, тада се једна од њих представља вектором $(1, 0)$, а друга $(0, 1)$.

Овакав приступ не уводи никакав поредак међу векторима јер они чине ортонормирану базу. Међутим, имплицитно је индукована линеарна зависност атрибута, јер важи $\sum_i v_i = 1$. Због тога се један од вектора мења нула вектором.

Применом описаног поступка, добијамо 40 нумеричких атрибута. Њихов број је очигледно превелик и да би се убрзао процес обучавања регресионог модела неке је потребно одстранити. Зато се на основу тренинг скупа, обучава неколико модела који могу проценити релевантност атрибута (попут насумичне шуме одлучивања или метода потпорних вектора). Уколико неки од тих модела одређени атрибут прогласи релевантним, додељује му се један глас. Бирају се сви атрибут са бар 3 гласа. На тај начин смо број атрибута више него преполовили. Напомињемо да су недостајуће вредности замењене медијаном. Изгласане атрибуте можемо видети у табели...

Потом се врши обучавање логистичке регресије. Како је потребно контролисати прилагођеност подацима, неопходно је одредити регуларизациони параметар регресионог модела. Како поменути параметар може узети вредност из непребројивог скупа, потребно је вршити узорковање или дискретизацију простора претраге. Опредељујемо се за први приступ и користимо `BayesSearchCV`. Претпоставља се \log -равномерна расподела параметра (у пакету `sklearn` се зове C). То значи да ће се параметар C узорковати из поменуте расподеле и чувати онај који је најбољи у односу

на задату метрику. Изводи се 300 симулација над искључиво тренинг скупом над којим се врши унакрсна валидација ради мерења ваљаности тако изабраног параметра. Бира се параметар за који модел логистичке регресије има највећу површину испод *ROC* криве. Водимо рачуна да се врши стратификована унакрсна валидација јер су класе неравномерно заступљене. Унакрсну валидацију лако изводимо употребом **StratifiedKFold** за $k = 5$.

Коначно, тако оптимизован **LogisticRegression** модел се тренира над тренинг скупом кроз 100 итерација или до конвергенције. Како постоји изражена неуравнотеженост класа, потребно је форсирати њихово балансирање у регресионом моделу, што на срећу, постоји као могућност у поменутом пакету

3.2 Имплицитно одстрањивање атрибута без увођења додатних атрибута

У претходно описаном поступку, релевантност атрибута је диктирана као хиперпараметар. Као алтернативан приступ, применићемо метод *weight of evidence* који не захтева елиминацију недостајућих вредности, експлицитно мењање категоричких атрибута нумеричким као ни експлицитно бирање релевантних атрибута.

Основна идеја је подела атрибута у дискретне скупове којима се придружује вредност на основу односа заступљености циљне променљиве у њима. Такву вредност зовемо *WOE* вредност. Затим се она додатно користи за пондерисање релативне заступљености позитивне класе чиме се добија *информациона вредност* (енг. *IV*) на основу које се може проценити релевантност атрибута.

Дајемо формално извођење поменутог метода. Нека је атрибут X_j подељен на строге подскупове B_1, \dots, B_n и нека је Y случајна величина циља. Вредност

$$\log \frac{P(X_j \in B_i | Y = 1)}{P(X_j \in B_i | Y = 0)}$$

означавамо са $WOE_{i,j}$. Можемо је оценити директно из узорка. Потом, информациону вредност атрибута X_j дефинишемо као

$$IV_j = \sum_{i=1}^n (P(X_j \in B_i | Y = 1) - P(X_j \in B_i | Y = 0)) \times WOE_{i,j}.$$

Број група се унапред фиксира и најчешће износи 20. Групе који имају сличну *WOE* вредност се стапају у једну.

Напомињемо да се приликом обучавања модела користи искључиво *WOE* вредност, док је информациона вредност коришћена искључиво за приказ релевантности атрибута.

На тај начин се целокупан поступак у великој мери аутоматизује, а најбитније, готово у потпуности уклања пристраност људске одлуке у коначном избору. Овај метод носи назив *WOEEncoder* у пакету *sklearn*¹.

У табели 1 је дат приказ првих 10 најважнијих атрибута према информационој вредности. На основу информационих вредности закључујемо да првих шест атрибута има умерену предиктивну моћ, док је код осталих она мање изражена. Ни један атрибут нема превелику предиктивну моћ ($IV > 0.5$) што би указивало да аномалију у подацима.

атрибут	IV_j
YearsAtCompany	0.284
TotalWorkingYears	0.263
MaritalStatus_Single	0.247
Age	0.237
YearsWithCurrManager	0.168
EnvironmentSatisfaction	0.123
BusinessTravel_Travel_Frequently	0.095
JobSatisfaction	0.083
MaritalStatus_Divorced	0.069
WorkLifeBalance	0.055

Табела 1: резултати на тест скупу

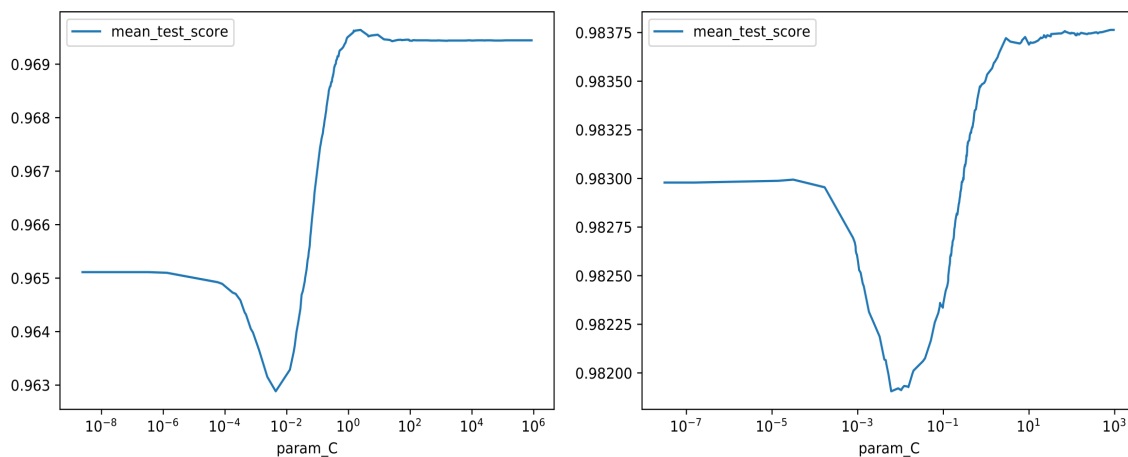
Након трансформације атрибута поменути поступком, врши се обучавање регресионог модела на већ описан начин.

На слици 2 се може видети како параметар C утиче на површину испод *ROC* криве.

4 Резултати

У табели 2 се може видети да модел има одзив 0.72 на позитивној класи и F_1 меру 0.81, за случај поделе скупа података 80/10/10. На тест скупу од 441 инстанци, модел постиже укупну прецизност 94%.

¹Како се на овај пакет често реферише, убудуће то неће бити навођено



Слика 2: вредности регуларизационог параметра применом Бајесове пре-траге. (лево) већи тренинг скуп (десно) мањи тренинг скуп

	прецизност	одзив	F_1	носач
0	0.93	0.99	0.96	360
1	0.92	0.72	0.81	81

Табела 2: резултати на тест скупу

5 Закључак

На основу приказаних резултата, имплицитни приступ се показао као бољи. Исход је очекиван имајући у виду напреднији начин одређивања релевантности и коришћења те информације у оквиру самог модела. Чак и ако експлицитни приступ обезбеђује сличне карактеристике, моделу се ни једног тренутка не даје информација о важности атрибута већ се то мора закључити на основу сирових података.

Постигнути су резултати који имају одличан одзив и F_1 меру над позитивном класом, што је и био циљ. За очекивати је да би успешност модела била већа када би скуп података био већи.

6 Имплементација

Сав код се може наћи на страници GitHub [репозиторијума](#).