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.

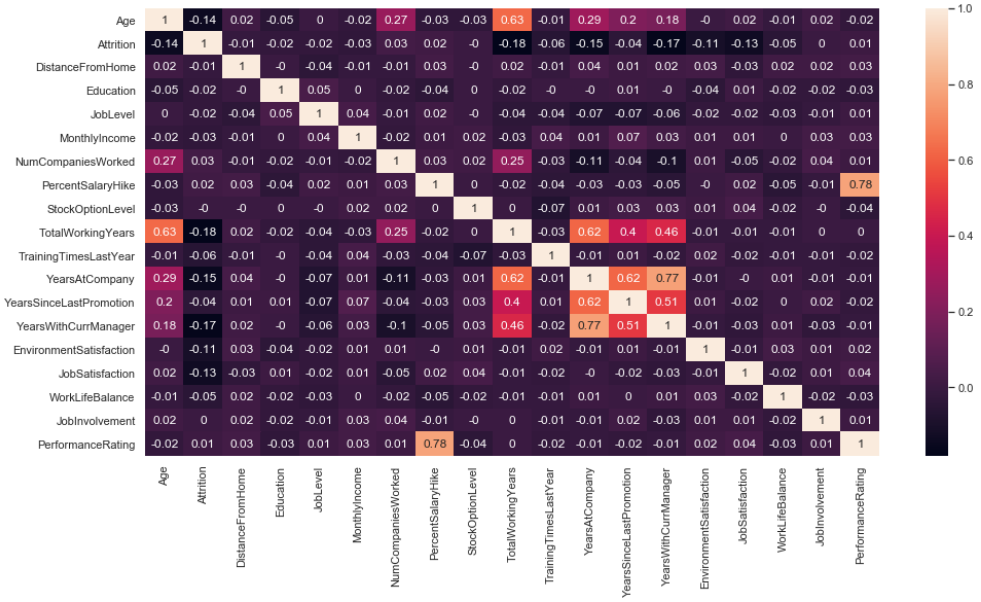
2 Подаци

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 категорички 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 релеватних 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 категоричке атрибуте, потребно је извршити њихову трансформацију. Зато ћемо описати одабир својстава и начин трансформисања категоричих атрибута.

Оба поменута проблема се могу решавати независно. Међутим, 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 одређује WOE вредност на основу односа вредности циљне променљиве у њима. Што је тај однос мањи то ће и WOE вредност бити мања, док је позитивна ако је однос у корист позитивног исхода, а негативна иначе. Затим се тако добијена вредност 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 кластере $B_1, ..., B_n$ и нека је Y случајна величина циља. Вредност

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

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

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

Број кластера се унапред фиксира и најчешће износи 20. Кластери који имају сличну WOE вредност се групишу у један.

На тај начин се целокупан поступак у великој мери аутоматизује, а најбитније, готово у потпуности уклања пристраност људске одлуке и коначном избору. Овај метод носи назив `WOEEncoder` у пакету `sklearn`[1]. У табели 1 је дат приказ првих 10 најважнијих атрибута према информационој вредности.

Важан параметар логистичке регресије је регуларизациони параметар (параметар C). Како га је готово немогуће одредити директно на основу података, потребно је извршити претрагу комбинаторног простора. У ту сврху се користи Бајесова претрага са унакрсном валидацијом (`BayesSearchCV`). Претпоставља се log-равномерна расподела параметра C. То значи да ће се параметар C узорковати из поменуте расподеле и

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

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

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 $k = 5$.

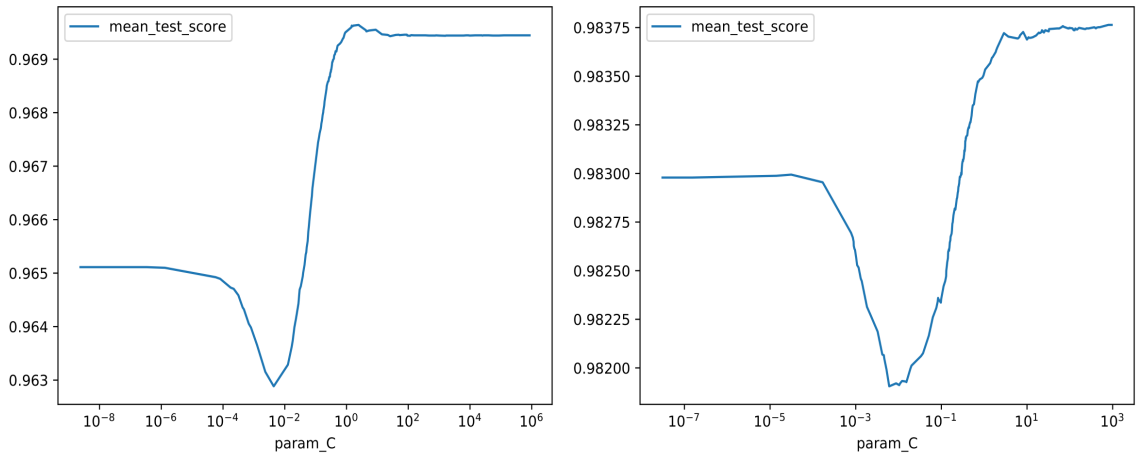
Параметар C, као што је и очекивано, показује завиност од величине података. Што је тренинг скуп већи, то је моделу исплативије да се што је више могуће прилагоди тренинг подацима, те је и регуларизација слабија, односно C је веће. Важи и обрат, мањи тренинг скуп повлачи јачу регуларизацију због мањег обима узорка.

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

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.

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%.



Слика 2: вредности регуларизационог параметра применом Бајесове претраге. (*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

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 кôд се може наћи на страници GitHub репозиторијума.