Универзитет у Београду

Електротехнички факултет



Модел за предикцију плате на огласима за посао

Извештај о домаћем задатку

|  |
| --- |
| Студент: Михајло Стојковић 2023/3137 |
|  |

Београд, Јануар 2023.

Садржај

[Садржај i](#_Toc125222899)

[1. Увод 1](#_Toc125222900)

[2. Анализа података 2](#_Toc125222901)

[2.1. Скуп за тренирање 2](#_Toc125222902)

[2.2. Преглед одређених улазних атрибута 4](#_Toc125222903)

[2.3. Анализа излазне вредности 5](#_Toc125222904)

[3. Одабир модела 8](#_Toc125222905)

[3.1. Линеарна Регресија 8](#_Toc125222906)

[3.2. Random Forest Regression 8](#_Toc125222907)

[3.3. XGBoost Regression (eXtreme Gradient Boosting) 9](#_Toc125222908)

[4. Закључак 10](#_Toc125222909)

1. Увод

Циљ овог задатка је генерисање модела за предвиђање плате на огласима за посао у Уједињеном Краљевству анализирајући њихов садржај.

Скуп података дат у изазову чини велики број стварних огласа (~245.000). Сваки оглас је распарчан у низ поља са релевантним информацијама извученим из самог огласа. То су следећа поља:

* ID – Јединствени идентификатор огласа за посао
* Title – Наслов огласа за посао. Уобичајено, ово је титула посла или сл.
* FullDescription – Целокупни текст огласа. У тексту огласа су маскирани подаци о плати.
* LocationRaw – Текстуална репрезентација локације директно извучена из огласа.
* LocationNormalized – Нормализована локација користећи стабло локација Adzuna-е. Ова нормализација није савршена.
* ContractType – full\_time или part\_time радно време интерпретирано из огласа помоћу Adzuna-е.
* ContractTime – permanent или contract. Означава да ли је посао на неодређено или одређено интерпретирано из огласа помоћу Adzuna-е.
* Company – Назив послодавца из огласа.
* Category – Којој од стандардних 30 категорија припада посао из датог огласа.
* SalaryRaw – Текстуална репрезентација плате директно извучена из огласа.
* SalaryNormalized – Нормализована годишња плата интерпретирана помоћу Adzuna-е. Ово је увек један број базиран као средња вредност опсега плате из огласа. Ово је вредност коју предвиђамо.
* SourceName – Назив сајта или оглашивача од ког је преузет оглас.

Сви подаци су стварни, живи подаци који се користе у огласима за посао тако да су очигледно подложни великој количини буке из стварног света, као што су огласи који нису у Великој Британији, плате које су погрешно наведене, поља која су погрешно нормализована, дупликати огласа итд.

За имплементацију модела је коришћен програмски језик *Python*.

1. Анализа података

У овом поглављу ће бити представљена детаљна анализа улазних података из пруженог *Train\_data\_rev1.csv* фајла. Разматране ће бити улазни подаци и како утичу на излазни податак.

* 1. Скуп за тренирање

Graphical user interface

Description automatically generatedНа следећој слици можемо видети првих и последњих 5 редова датог скупа за тренирање модела.

У наставку ће бити приказане још неке релевантне информације о датом скупу података.

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Може се приметити да одређене колоне (ContractType, ContractTime, Company) имају доста празних вредности што нам даје до знања да ће имати нешто мањи утицај на предвиђање модела. Свакако, приликом сређивања података, за сваку колону која је имала празна поља је дефинисана нова *boolean* колона која је означавала да ли оглас има или нема дефинисано дато поље (нпр. ContractTypeIsMissing).

Како је ово регресивни проблем јер модел треба да предвиђа реалан број, у наставку ће се посебно размотрити одређене улазне колоне.

* 1. Преглед одређених улазних атрибута

Chart, bar chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedУ овом поглављу ће бити размотрени одређени улазни атрибути који су од интереса из скупа података.

Из датих графика се може сагледати квалитет улазних података. Како су ово подаци из реалног света, види се да нису савршени. На графицима са претходне стране су приказане колоне које имају релативно мали број могућих вредности. Оно што се може приметити јесте да у оба случаја подаци нагињу ка некој вредности или неким вредностима које се појављују знатно више од осталих. У случају локације, најчешћа локација у огласима је *UK* и *London* што представља проблем за прецизност модела јер неће имати конзистентно висок ниво информација о огласима у сваком граду. Са друге стране, приликом анализе броја огласа по категорији посла, може се приметити да се *IT, Engineering, Accounting & Finance* и *Healthcare & Nursing* појављују знатно чешће од осталих. Иако је ситуација овде нешто боља и даље се одређене категорије јако ретко појављују у скупу података што ће имати утицај на прецизност модела.

* 1. Анализа излазне вредности

На наредној слици су приказане плате из улазног скупа података расејане на дијаграму:

Chart

Description automatically generatedМоже се приметити да постоји мали број екстремума плата (изнад 100.000 и испод 10.000) па је одлучено да се одсеку горњих и доњих 1% вредности. Након тог поступка распоред тачака је нешто линеарнији што ће побољшати прецизност модела. На наредној слици је приказан дијаграм након одсецања.

Chart, histogram

Description automatically generatedChart

Description automatically generatedНа следећем хистограму је приказан број плата у одређеним опсезима:

Постоји доста већи број огласа са платама мањим од 50.000 док оних изнад има нешто мање. Поново, ово је знак да подаци нису савршени те да ће модел имати проблема са предвиђањем високих плата како нема равномерно распоређене улазне вредности за плату.

Chart, waterfall chart

Description automatically generatedНа крају, од користи је и анализирати топлотну мапу корелација између колона. Таква мапа је приказана испод:

Овде је од интереса генерално приметити зависности између колона али оно што се може одмах приметити јесте да свака колона има релативно ниску корелацију са *SalaryNormalized* док једина пристојна корелација је са *SalaryRaw* колоном.

1. Одабир модела

У овоме поглављу биће разматрани различити регресивни модели и њихова прецизност над датим проблемом.

За поређење су одабрани Линеарна Регресија, *Random Forest Regression* и *XGBoost Regression* као једни од најзаступљенијих регресивних алгоритама. За поделу података на део за тренирање и део за тестирање се користила *train\_test\_split* метода из *sklearn* библиотеке подешена тако да се 25% података користи за тестирање а остатак за тренирање. Како би регресивни модели функционисали, потребно је претворити све текстуалне вредности у јединствене идентификаторе. Ово је постигнуто уз помоћ *LabelEncoder* библиотечке функције.

* 1. Линеарна Регресија

Линеарна регресија је статистичка метода која се користи за моделирање односа између зависне променљиве и једне или више независних променљивих прилагођавањем линеарне једначине посматраним подацима. Циљ линеарне регресије је да се направи модел који може предвидети вредност зависне променљиве на основу вредности независних променљивих. Линеарна регресија се често користи за предвиђање и анализу података. Линеарна регресија претпоставља да постоји линеарна веза између независних и зависних варијабли, да су грешке нормално распоређене и да имају константну варијансу. Модел проналази линију која најбоље одговара тако што минимизира збир квадрата разлика између предвиђених и стварних вредности.

Модели линеарне регресије су релативно једноставни и пружају математичку формулу лаку за интерпретацију која може да генерише предвиђања. Линеарна регресија се може применити на различите области. Пошто је линеарна регресија дуго коришћена статистичка процедура, својства њеног модела су добро схваћена и могу се врло брзо обучити.

На расејаном дијаграму плата се може приметити да плате не прате савршено праву линију те се може предпоставити да овај модел неће дати идеалне резултате. Као оптимизација линеарне регресије се користила полиномна регресија другог степена. Ово је дуплирало прецизност модела али је и даље прецизност значајно гора од других модела.

* 1. Random Forest Regression

*Random Forest* је ансамбл стабала одлучивања. То значи да многа стабла, изграђена на одређени „случајни“ начин, формирају *Random Forest*. Свако стабло се креира из различитог узорка редова и на сваком чвору се бира другачији узорак карактеристика за раздвајање. Свако од стабала затим даје своје индивидуално предвиђање. Ова предвиђања се затим усредњују да би се произвео један резултат. Усредњавање чини *Random Forest* бољим од једног стабла одлучивања, па стога побољшава тачност и смањује прекомерно прилагођавање.

Стабла одлучивања су одлична за добијање нелинеарних односа између улазних карактеристика и циљне променљиве. Унутрашњи рад стабла одлучивања може се посматрати као гомила if-else условa.

За овај проблем, коришћен је класичан *Random Forest* као и побољшана верзија са тјунираним параметрима. Овај алгоритам се показао најпрецизнијим над пруженим скупом података.

* 1. XGBoost Regression (eXtreme Gradient Boosting)

*XGBoost* је *open-source* библиотека за *gradient boosting* која се користи за проблеме машинског учења као што су регресија и класификација. XGBoost је имплементација *gradient boosting*-а која користи стабла одлучивања као слабог ученика.

У *XGBoost* регресији, циљ је да се предвиди континуирана варијабла. Функционише тако што гради ансамбл стабала одлучивања корак по корак, где се свако дрво уклапа у негативни градијент функције губитка. Ово омогућава алгоритму да се фокусира на грешке претходних стабала, а нова стабла ће затим исправити те грешке. Коначно предвиђање је пондерисани просек предвиђања из свих стабала.

*XGBoost* се широко користи у пракси јер је ефикасан, има добре перформансе и једноставан је за коришћење. То је један од најбољих алгоритама који се користи на *Kaggle* такмичењима, где често побеђује.

Иако он генерално даје боље резултате од *Random Forest* модела, у овом случају је био врло сличан али не и бољи од њега.

1. Закључак

На наредној слици се могу видети упоређени резултати тестираних алгоритама по параметрима средње апсолутне грешке, корена средње квадратне грешке као и коефицијент детерминације R2:

Graphical user interface, text

Description automatically generated

Ови резултати минимално варирају сваком итерацијом тренирања модела али пошто се располаже великом количином улазних података, варијација је минимална. Највише варира тјунирани *Random Forest* како он бира рандом вредности параметара. Посматрајући резултате, може се закључити да је *Random Forest* показао највећу прецизност док је близу иза њега *XGBoost.* Тјунирани *Random Forest* би можда могао и да надмаши резултате класичног уз довољну процесорску моћ. Свакако, коефицијент детерминације од ~0.93 представља врло задовољавајући резултат узимајући у обзир шум података из реалног света.