Fake Job

Maialen Ruiz
Datu meatzaritza
UPV/EHU Unibertsitatea

eman ta zabal zazu

Universidad del País Vasco — Euskal Herriko Unibertsitatea

Introduction

Proiektu honetan, kaggle-n aurkitutako fake job postings dataset-a erabili da. Datu hauek bi klase dituzte, fake (0) edo real (1) eta 17880 instantzia kopuru biltzen ditu. Originalki datuak atributu ugari zituen, baina ataza honetarako denak elkartu ditut. Testu hauen klasea iragartzeko errepresentazio eta klasifikatzaile desberdinak erabili ditut: TFIDF, DOC2VEC, Logistic Regression eta MLP.

Text Representation

Datuei aplikatutako aurreprozesamendua:
- 18 atributuak elkartu
-StopWords (english)
-Lematizatu

	post	Fraudulent
0	Marketing Intern,US, NY, New York,Marketing,We...	0
1	Customer Service - Cloud Video Production,NZ, ...	0
2	Commissioning Machinery Assistant (CMA),US, IA...	0
3	Account Executive - Washington DC,US, DC, Wash...	0
4	Bill Review Manager,US, FL, Fort Worth,SpotSou...	0
...	...	...

TFIDF (sklearn liburutegia)
min_df
1
3
DOC2VEC (gensim)
alpha=0.002

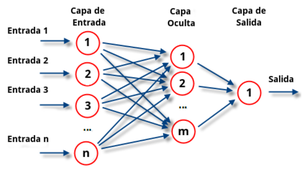
VectorSize	mincount
150	1
150	2
300	1
300	2

Logistic regression

Estatistikan, erregresio logistikoa edo logit eredua gertakizun baten probabilitatea aurresateko erabiltzen den erregresio-teknika bat da, aldagai independente zenbaitetan oinarrituta kurba logistiko bat egokituz.

MLP

Geruza anitzeko pertzepzioa geruza ugariz osatutako sare neuronal artifiziala da, eta linealki bereiz ezin daitezkeen arazoak konpontzeko gaitasuna du.

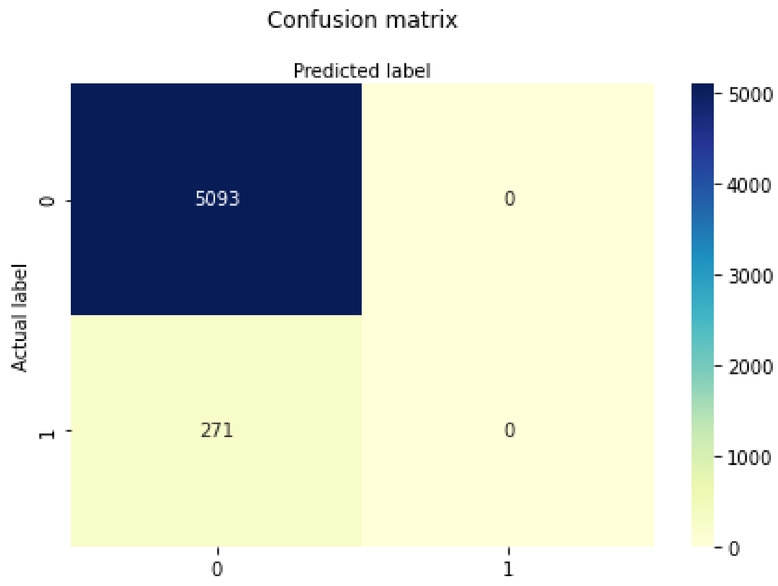


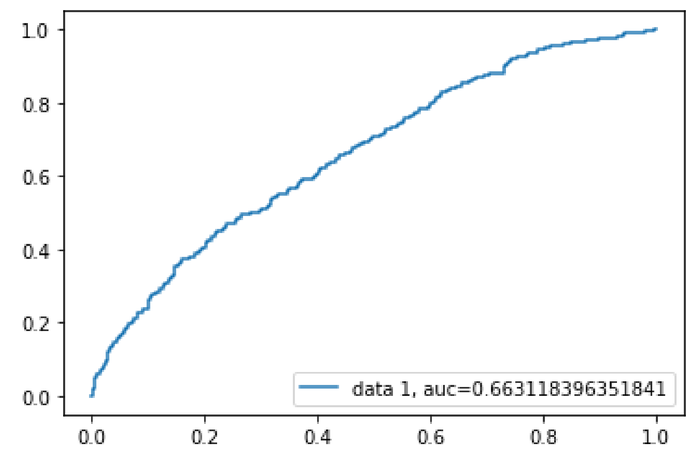
EMAITZAK

F-SCORE	Doc2Vec				TFIDF	
	vectorSize=300 mincount=1	vectorSize=300 mincount=2	vectorSize=150 mincount=1	vectorSize=150 mincount=2	min_df=1	mindf=3
LogisticRegression	0,93	0,93	0,93	0.925	0,95	0,96
MLP	0,92	0.93	0.92	0.92	0.95	0,96

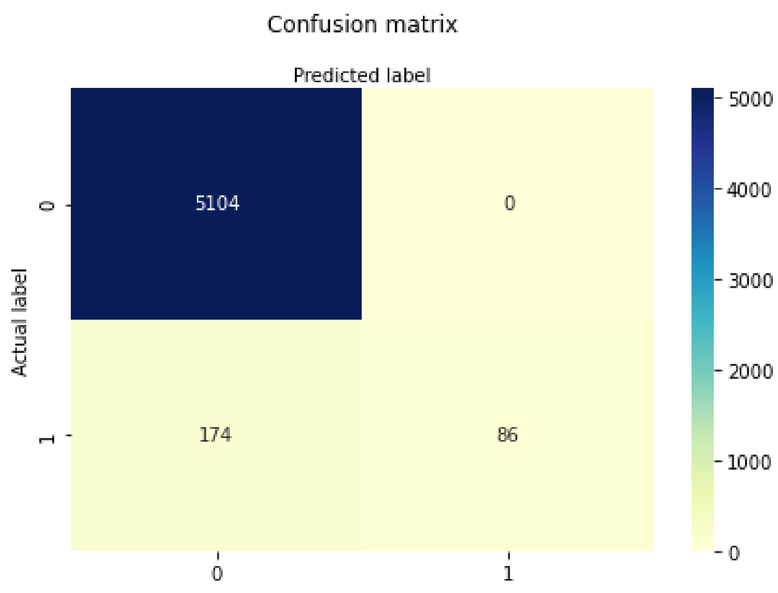
Logistic Regression - Doc2Vec/Tfldf

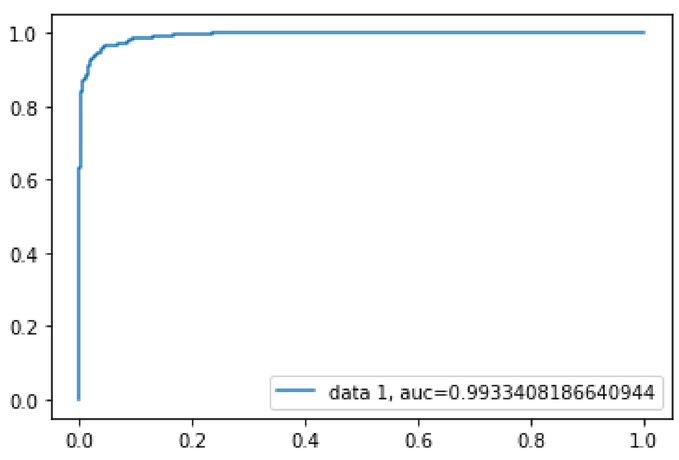
Doc2Vec errepresentazioa aplikatuz lortu dudan nahaste matrizea eta ROC grafikoa:





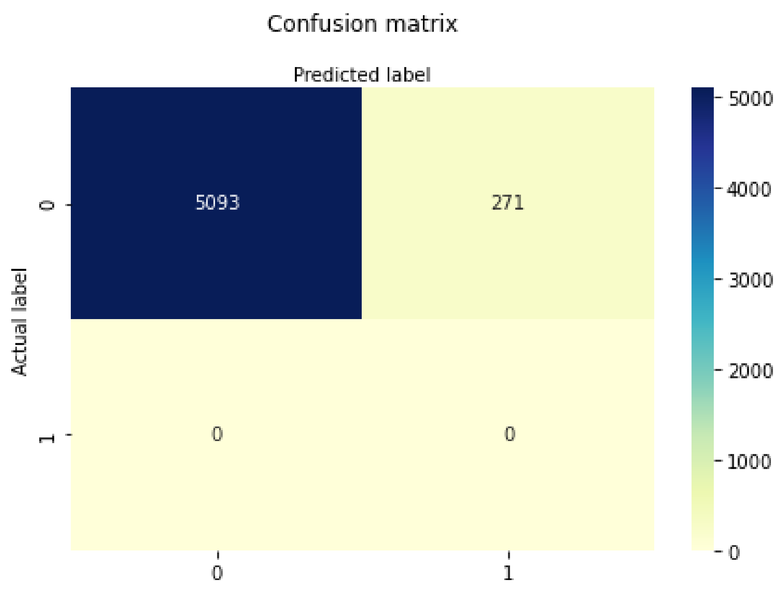
TFIDF errepresentazioa aplikatuz lortu dudan nahaste matrizea eta ROC grafikoa:

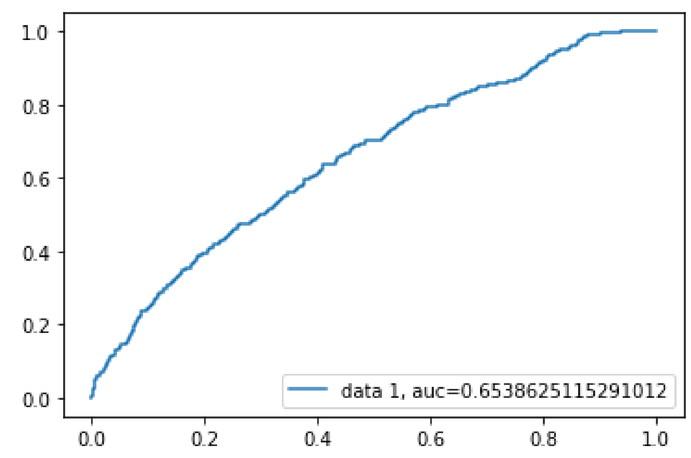




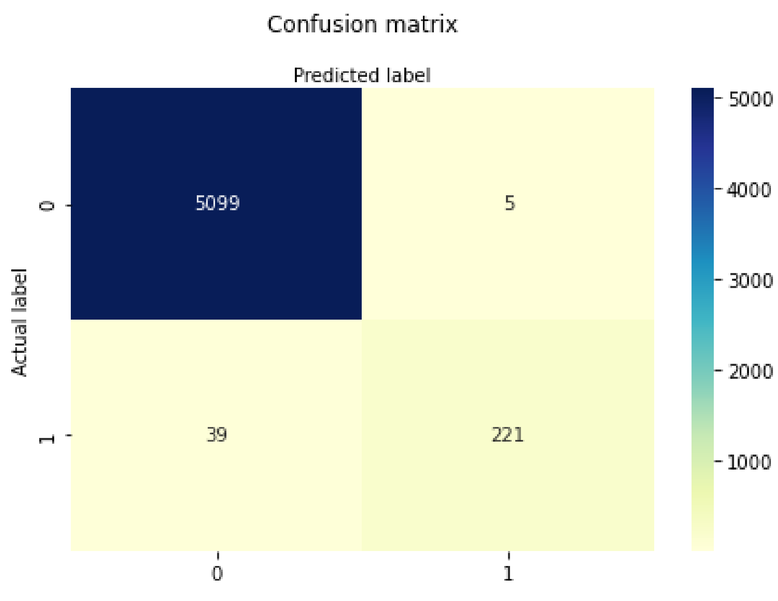
MLP - Doc2Vec/Tfldf

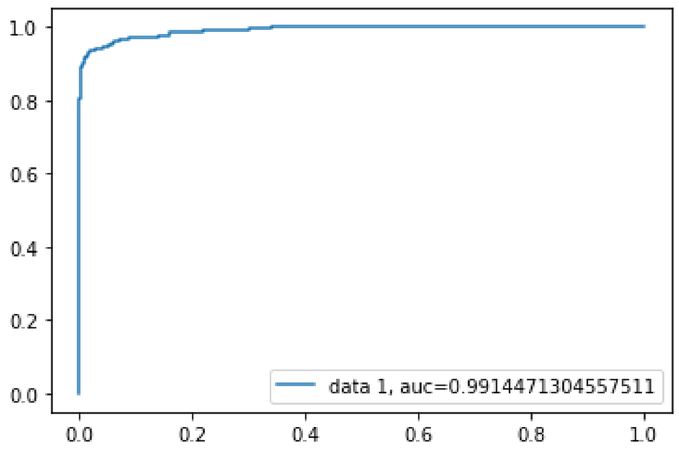
Doc2Vec errepresentazioa aplikatuz lortu dudan nahaste matrizea eta ROC grafikoa:





TFIDF errepresentazioa aplikatuz lortu dudan nahaste matrizea eta ROC grafikoa:





Ondorioak

Emaitza guztiak aztertu ondoren ikusi dezakegu ez daudela kriston desberdintasunik. F-score-an zentratzen bagara TFIDF-MLP modeloa onena dela esan dezakegu. Datu sorta ez dago ondo orekatuta, grafikoan ikusten dugun bezala %5-a fake klasekoa da eta %95 real klasekoa. Beraz, accuracy eta f-score onak lortu arren baliteke ez egotea guztiz ondo entrenatuta fake lan postuak iragartzeko. Etorkizunera begiratuz, datu sorta handitzea gustatuko litzaidake egoera guztiak ondo aztertu ahal izateko eta errepresentazio eta klasifikatzaile ereduen aldetik, beste parametrizazio sakonago bat egitea.