Predicció de temps en trobar feina d'una persona amb sordesa

Pere Girona Campi
https://github.com/perot84
pere.girona@onmail.com

Abstracte

El projecte intenta fer una predicció de quan triga a trobar feina un usuari del servei d'inserció laboral d'una entitat que treballa amb persones. S'ha treballat amb diferents dades de bases extretes del CRM de la pròpia entitat i el s'ha obtingut un resultat molt limitat, degut en part a la poca quantitat i qualitat de les dades. Cal tenir en compte que la recollida de totes les dades és manual, la qual cosa pot provocar diferents criteris a l'hora d'introduir-les (tant com es recullen, com si es recullen o no).

1. Introducció

Per a crear el conjunt de dades s'han extret diferents bases de dades a partir del CRM de l'entitat. Ens hem trobat amb un conjunt de dades força limitat en nombre i varis problemes en la consistència de dades, en la introducció d'aquestes i amb un nombre força elevat de valors faltants. Hem aprofitat aquesta situació per a ampliar els objectius prèviament descrits. Hem decidit crear una base de dades on unir totes les dades pertanyent a cada usuari, així com introduir i seleccionar característiques no recollides anteriorment, fer un recull de propostes de millora la de recollida i gestió de les dades, així com argumentar la necessitat d'una bona recollida de dades a partir de la utilitat que pot donar a les persones que les introdueixen. En aquest últim apartat cal destacar que qui fa més ús d'aquestes dades no són les persones tècniques que les introdueixen, si no que moltes vegades és personal tècnic superior, per aquest motiu hem pensat en aquest objectiu.

Entitat social d'on extraiem les dades i amb la qui treballem:

L'Associació Catalana per a la Promoció de les Persones amb Sordesa (ACAPPS), creada l'any 1992, és una entitat sense ànim de lucre declarada d'Utilitat Pública que té com a finalitat representar i defensar els interessos globals de les persones amb discapacitat auditiva i de les seves famílies. ACAPPS té com a àmbit territorial Catalunya. Les tres associacions (ACAPPS Barcelona, ACAPPS Lleida i ACAPPS Vallès) formen part de la Federació d'Associacions Catalanes de Pares i Persones amb Sordesa (ACAPPS).

La missió d'ACAPPS és representar i defensar els drets i interessos globals de les persones amb sordesa i de les seves famílies, a tots els nivells, davant la societat, administracions públiques i altres institucions, integrant i impulsant amb aquesta finalitat l'acció de les famílies, o representants legals, i de les persones amb sordesa.

2. Metodologia

En aquest projecte el preprocessat ha tingut un paper molt important, ja que hem hagut d'utilitzar 7 bases de dades diferents amb múltiples errors, migracions incompletes, valors faltants, duplicats d'ids...

L'objectiu del preprocessat ha estat mantenir el màxim nombre de característiques per a millorar les mètriques alhora que s'han creat noves columnes. Destacar que s'ha creat la la columna de temps que triga la persona a trobar feina, la qual ha estat l'objectiu a preveure.

Hem provat diferents models regressius per a cercar el temps que triga un usuari a trobar feina en què el millor resultat ens l'ha donat KNeighborsRegressor.

Hem usat les mètriques R2, MSE i MAE. Degut a que han sortit uns resultats molt baixos ens hem guiat per la MAE, el qual ens indica l'error absolut. En aquest cas explica quan dies d'error té el model, la qual cosa també ens permetia entendre i transmetre millor el significat al client.

3. Resultats

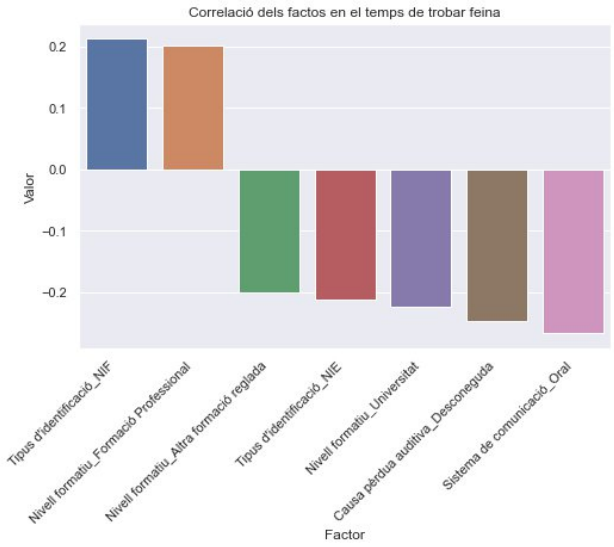
	Model	RMSE	MAE	R2
0	ElasticNet	128.494145	109.729904	-0.119131
1	RandomForestRegressor	133.981949	110.579821	-0.216765
2	AdaBoostRegressor	133.471857	104.010536	-0.207518
3	SVR	122.299303	102.137339	-0.013823
4	SVR	122.985756	102.088671	-0.025236
5	KNeighborsRegressor	139.000653	109.276316	-0.309628

El millor resultat l'hem obtingut amb *SVR* amb un resultat en el MAE de 102.

El model SVR és un algoritme d'aprenentatge automàtic supervisat dirigit a problemes de regressió. Treballa a partir de vectors i està especialment indicat per a treballar amb dades no lineals. Aquest resultat implica que el model té un error de 102 dies en la seva predicció.

La previsió no ha obtingut un resultat massa bo. Cal destacar que hem treballat amb poques entrades (77) i eliminant algunes variables per no tenir suficients dades o ser de poca qualitat. Creiem que si es millora la qualitat i quantitat de dades es podrien millorar de manera significativa els resultats.

Per altra banda hem recollit els valors de correlació entre les característiques i la variable amb els següents resultats:



Com podem veure el fet que la persona es comuniqui oralment i que tingui una formació universitària seran dos dels factors més importants per a que el temps per a trobar feina sigui menor. Per altra banda la formació professional i que la persona tingui NIF seran factors que propiciaran un augment en el temps a trobar feina. Cal donar una importància relativa a les dades, ja que disposem de poques dades, per la qual cosa aquest correlació tindrà variacions importants a partir de l'entrada de noves dades.

4. Conclusions

Tot el procés del projecte ha estat centrat en oferir i donar utilitat a l'aprenentatge adquirit a una entitat social, sense contacte previ en la ciència de dades.

Encara que el resultat de la predicció ha estat limitat s'han pogut aportar diferents coneixements a l'entitat:

Millores de recollida de les dades

Millores en la gestió de les dades (CRM)

Coneixement i utilitat de la ciència de dades per a extraure més profit a les dades recollides.

Donar més valor a la introducció de les dades, la qual cosa pot repercutir en una millora de la quantitat i qualitat d'aquestes.

Per altra banda hem comprovat com és d'important explicar prèviament i de la millor manera les possibilitats del Machine Learning, ja que a mesura que han avançat les reunions hem pogut recollir més demandes del client i donar més possibilitats a la feina realitzada.

Per a que la utilitat fos més gran caldria crear una API per tal d'automatitzar la recollida de dades i la creació de les base de dades.