Exemples de projectes en l'àmbit de la ciència de dades

PID_00261828

Marçal Mora Cantallops

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

Marçal Mora Cantallops

Enginyer industrial i enginyer informàtic per la UPC, màster en Data Science per la UAH i doctorand en Comunicació, Informació i Tecnologia de la Societat en Xarxa per la mateixa universitat. Investigador en l'àmbit dels game studies, la ciència de dades i, en particular, l'anàlisi de xarxes socials; està interessat en l'ús d'aquestes tècniques per a l'extracció de coneixement i informació. Ha treballat en la creació i optimització de models estadístics per a logística i planificació de la demanda i actualment participa en diversos projectes relacionats amb l'estadística i la ciència de dades.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Josep Maria Marco (2019)

Índex

Introducció

Després d'aquest breu però intens recorregut per la ciència de dades i el seu entorn, és el moment de presentar alguns exemples d'aplicacions, que serviran per entreveure les possibilitats que aporta aquesta disciplina. Els projectes de ciència de dades, com veurem, poden ser molt diversos, des de tasques humanitàries (en les quals les Nacions Unides ha posat especial atenció) fins a situacions empresarials, passant per importants aplicacions sanitàries o classificació d'espècies, entre d'altres.

En aquest mòdul exposarem alguns d'aquests casos però ho farem des d'una perspectiva general. El més important és entendre quin tipus de problemes es planteja resoldre la ciència de dades, els processos que s'apliquen en la seva resolució i la forma que tenen les conclusions o les observacions que se'n poden extreure.

En els tres primers apartats d'aquest mòdul tractarem un total de tres grans casos (tot i que el primer es divideixi en tres exemples més concrets) en què la ciència de dades i les dades massives tenen un paper determinant. Anirem des d'una perspectiva més general, a vista d'ocell, fins a un cas complet, que permetrà a l'estudiant veure el flux d'un petit projecte des del principi fins al final.

L'estructura del mòdul és la següent:

1) En primer lloc, veurem alguns dels projectes més destacats dels disponibles a la pàgina del projecte Global Pulse de les Nacions Unides (https://www.unglobalpulse.org/projects).

2) En segon lloc, veurem una aplicació molt important: l'ús de la ciència de dades per al control d'epidèmies. Veurem, en particular, el cas de l'Ebola.

3) En tercer lloc, veurem un exemple que, tot i ser plantejat de manera molt didàctica, podria ser un cas d'activitat empresarial rellevant per a una empresa de transport urbà: l'anàlisi dels recorreguts dels taxis a Nova York.

4) Per acabar, repassarem les idees més importants del mòdul.

Aquest mòdul és, doncs, una oportunitat d'il·lustrar la ciència de dades mitjançant el que fa i, sobretot, el que pot fer.

1. Projectes de ciència de dades per al desenvolupament i l'acció humanitària

El projecte Global Pulse és una iniciativa d'innovació de les Nacions Unides que utilitza la ciència de dades per contribuir al desenvolupament sostenible. Aquest projecte consisteix en l'extracció i l'anàlisi responsable de dades massives amb l'objectiu d'aportar alguna cosa positiva a la humanitat. La seva missió és «accelerar el descobriment, desenvolupament i adopció de la innovació en dades massives per al desenvolupament sostenible i l'acció humanitària».

Aquesta iniciativa, que ja té uns anys, es va establir basant-se en la idea que les dades ens donen la possibilitat d'entendre millor la humanitat i els canvis en el seu benestar, així com d'obtenir una imatge en temps real sobre la resposta de la humanitat als canvis polítics, normatius i legals.

Així doncs, Global Pulse treballa per conscienciar els ciutadans de les possibilitats de la ciència de dades en el desenvolupament humà i l'acció humanitària, però també busca formar aliances entre l'entorn públic i privat per compartir informació, generar eines analítiques i metodologies que puguin ser utilitzades àmpliament, i estendre els avanços a tota la xarxa de les Nacions Unides.

Figura 1. Com la ciència de dades pot contribuir al desenvolupament sostenible.

Font: United Nations Global Pulse (https://www.unglobalpulse.org/about-newhttps://www.unglobalpulse.org/about-new)

Alguns dels àmbits en què s'agrupen els projectes (i que donen una idea clara del seu objectiu) són:

- pobresa i fam

- educació i treball de qualitat
- igualtat de gènere
- energia neta i acció contra el canvi climàtic
- ciutats sostenibles

El funcionament dels seus projectes és el següent: una sèrie de laboratoris d'innovació generen idees i coordinen projectes, associats amb experts de cada un dels àmbits, governs, acadèmics i el sector privat. Així, desenvolupen i investiguen diverses aproximacions per a l'aplicació de la informació digital (habitualment, en temps real o gairebé real) als reptes que presenta el segle XXI.

Com veurem en els projectes que analitzarem a continuació, el plantejament del procés és el següent:

1) Obtenir accés a les dades, eines i experiència necessàries per descobrir noves aplicacions.

2) Desenvolupar les eines, aplicacions i plataformes que puguin millorar la presa de decisions o l'avaluació de les mateixes.

3) Contribuir al desenvolupament de marcs regulatoris que assegurin l'ús ètic de les dades i la privacitat.

4) Involucrar actors clau (governs, empreses, ciutadans) en les prioritats d'innovació i proporcionar-los assistència en la implementació.

1.1. La deducció dels desplaçaments diaris dels habitants de Jakarta a partir de les dades de Twitter

Diuen que Jakarta, o la seva àrea metropolitana, té més de 30 milions d'habitants. A la ciutat, el sistema de transport public gestiona aproximadament 1,38 milions de desplaçaments diaris. Tot i que semblin molts (i que ho siguin), hem de pensar que ciutats com Barcelona o Madrid, amb àrees metropolitanes cinc vegades més petites, gestionen uns 4 milions de viatges diaris en transport públic, segons els seus respectius ajuntaments. El motiu és simple: aquestes dues ciutats tenen sistemes de metro centenaris, mentre que Jakarta va inaugurar la seva primera línia l'any 2013, en una ciutat caracteritzada pels embussos de trànsit i del transport rodat (habitualment, de dues rodes).

Tant el gran nombre d'habitants com l'estat de la infraestructura de transport (i la saturació de la carretera) converteixen els desplaçaments diaris en una queixa habitual dels residents. Per aquest motiu —i per la sostenibilitat de la ciutat— el govern local treballa per millorar la situació; un dels camins triats és seguir desenvolupant noves línies de metro. Però, com s'ha de decidir el traçat que han de tenir? L'oficina d'estadística d'Indonèsia va optar per la via habitual: una enquesta. Des del disseny de l'enquesta fins a l'obtenció dels

resultats va passar més d'un any. Seria possible estalviar temps, diners i esforços i obtenir resultats similars o igualment útils aplicant alguna tècnica de ciència de dades?

1.1.1. La recollida de dades

La importància de les dades per a la planificació urbanística no és única de Jakarta, totes les grans ciutats del món hi estan interessades. Potser la manera més fàcil de traçar els desplaçaments sigui utilitzant la informació que surt d'un dispositiu petit que gairebé tothom porta a la butxaca: el telèfon mòbil. La inclusió de GPS, sensors i plataformes socials dona molt de joc. A Indonèsia, i a Jakarta en particular, les xarxes socials estan molt esteses; l'any 2012, de fet, va ser nomenada com la capital mundial de Twitter, perquè era la ciutat amb més activitat a la plataforma.

És possible aprofitar la geolocalització dels missatges enviats a Twitter per dibuixar els patrons de desplaçament a l'àrea metropolitana de Jakarta? El projecte va consistir a extreure milions de missatges de la xarxa social i analitzar les relacions entre els deu districtes (o ciutats de l'àrea) principals. També és important, no obstant, tenir en compte que Twitter pot ser una representació esbiaixada de la realitat, ja que no tots els estrats de la població tenen la mateixa possibilitat d'accés a la xarxa social (penseu, per exemple, en la gent gran o les persones sense telèfon). Per aquest motiu és necessari calibrar el resultat final per afinar-lo.

1.1.2. L'anàlisi dels desplaçaments

Els investigadors van recollir totes les piulades que tenien dades de GPS a l'àrea metropolitana de Jakarta entre l'1 de gener de 2014 i el 30 de maig del mateix any, coincidint amb l'extracció de l'enquesta oficial.

Aleshores, per a cada usuari:

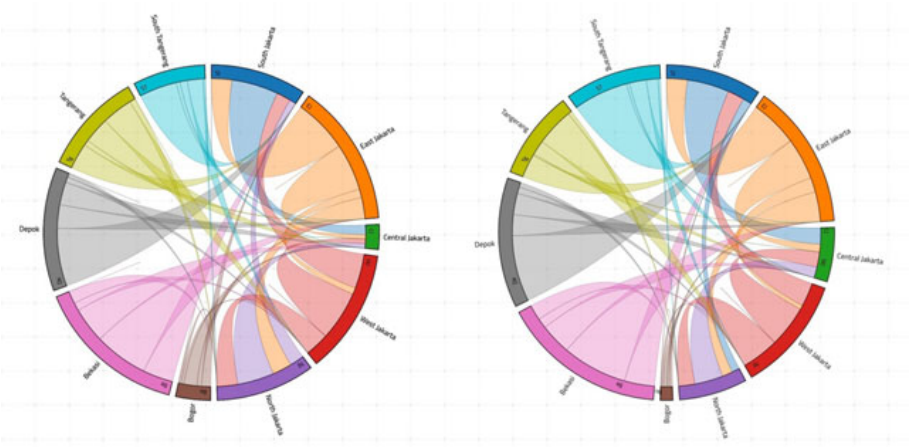
- L'origen s'estableix en el lloc des del qual escriu més missatges entre les nou del vespre i les set del matí.

- La destinació s'estableix en el lloc des del qual escriu més durant la setmana laboral, excloent-ne l'origen.

És important fixar-se en la manera de determinar l'origen i la destinació perquè no són assignacions directes, sinó interpretacions de patrons de dades. Que l'origen d'una persona sigui allà on sopa, dorm i es desperta, pot tenir sentit, de la mateixa manera que on escriu la majoria de missatges de dia (i entre setmana) pot ser a prop del seu lloc de destinació.

Amb aquest mètode es va obtenir informació de més de 300.000 usuaris únics (tot i que s'havien recollit dades d'aproximadament milió i mig d'usuaris). Per calibrar els resultats es va ponderar segons la població total de cada districte, per convertir les dades de Twitter en més proporcionals. A la figura 2 s'hi poden veure els resultats.

Figura 2. Comparativa entre l'enquesta i les dades de Twitter



A l'esquerra, els moviments de l'enquesta oficial i a la dreta, els extrets de Twitter Font: https://www.unglobalpulse.org/inferring-jakarta-commuting-statistics-twitter

Com es pot comprovar, els resultats obtinguts mitjançant l'anàlisi dels missatges a Twitter són molt semblants als obtinguts a partir de l'enquesta oficial; l'avantatge, no obstant, és que el monitoratge mitjançant Twitter pot ser pràcticament en viu i continu. A més a més, els resultats s'obtenen de manera molt més ràpida i sembla que no es perd qualitat. Algú podria argumentar, fins i tot, que els resultats potser siguin més realistes. Al cap i a la fi, qui no ha mentit alguna vegada en una enquesta?

1.2. L'ús de les dades massives per estudiar els patrons de rescat a la mar Mediterrània

Viatges fatals

L'informe complet, molt interessant per veure també les visualitzacions, està disponible a https://bit.ly/2xRXmmg.

Aquest projecte neix l'any 2017, després que les Nacions Unides declarés 2016 com l'any amb més morts de migrants la mar Mediterrània. L'objectiu era reduir la xifra de víctimes mortals. Però per aconseguir-ho feia falta visualitzar, primer, quin era el procés habitual de salvament que seguien els vaixells, majoritàriament d'ONG, per tal d'entendre les dificultats principals a les quals s'enfrontaven.

Els científics de dades van utilitzar les dades de localització proporcionades per un sistema anomenat AIS i que proporciona informació sobre la posició i la velocitat dels vaixells que envien aquestes dades (així com el seu identificador, curs i destinació, entre d'altres) de manera regular, habitualment cada dos minuts. És la informació que utilitzen les autoritats marítimes per, entre d'altres, evitar col·lisions entre naus.

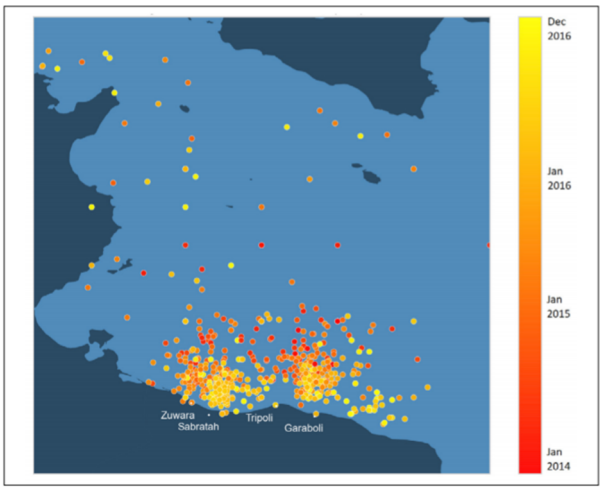
Amb aquesta informació és possible, per exemple, dibuixar visualitzacions com la de la figura 3, que més que un gràfic és una narrativa que explica una història. En aquest cas, es tracta d'una de les operacions de l'Aquarius (el vaixell de les ONG SOS Mediterraneé i Metges sense Fronteres) i el seu trajecte rescatant múltiples embarcacions a la deriva.

Figura 3. Un rescat seqüencial de més d'un naufragi

Font: https://publications.iom.int/system/files/pdf/fatal_journeys_volume_3_part_1.pdf

De la mateixa manera que els vaixells transmeten informació sobre la seva posició regularment, existeix un sistema d'auxili que serveix per notificar possibles problemes en una regió o, més important, per notificar una emergència als vaixells propers (lligant-los legalment a respondre, si és possible). Aquests avisos contenen el nombre de persones estimades a bord i la localització aproximada.

Figura 4. Rescats en casos de naufragi



De color vermell, els rescats més antics i de color groc, els més recents. Els rescats tenen lloc cada vegada més a prop de la costa. Font: https://publications.iom.int/system/files/pdf/fatal_journeys_volume_3_part_1.pdf

Una de les troballes més importants (o confirmació del que molts afirmaven basant-se en la seva observació) és que les crides de socors passaven cada vegada més a prop de la costa de Líbia (figura 4) i obligaven les operacions de rescat a entrar en zones en què no solien fer-ho. Així, per salvar el major nombre

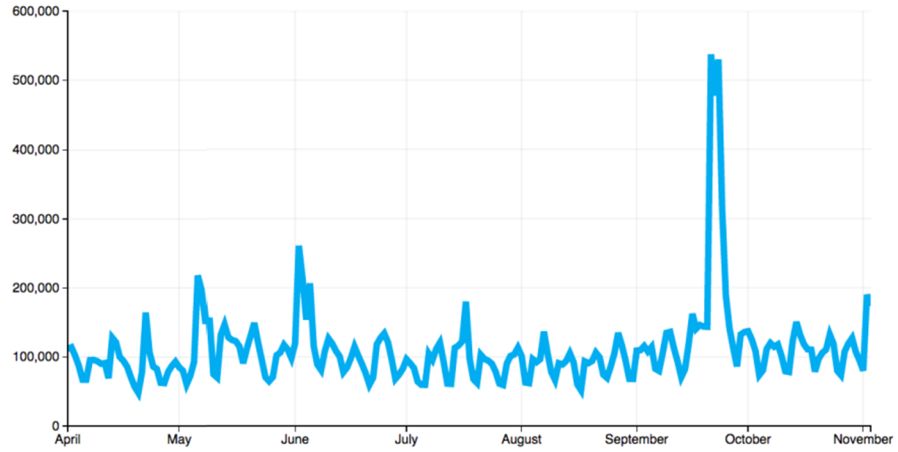
de persones, és necessari cobrir zones marítimes cada vegada més grans; si els efectius o vaixells de salvament són els mateixos o cada vegada menys, més persones acaben perdent la vida a la recerca d'un futur millor.

1.3. Twitter i la implicació global sobre el canvi climàtic

Una de les maneres de recollir l'opinió o l'activitat de la població sobre un tema concret és monitorar les xarxes socials. Moltes empreses ho fan per obtenir informació sobre els seus productes, per exemple. En aquest cas, l'objectiu de les Nacions Unides era generar un observatori en temps real sobre el discurs dels usuaris de Twitter en l'àmbit global (sí, a escala mundial) respecte al canvi climàtic abans, durant i després de la cimera del clima de 2014.

La idea és que Twitter pot actuar com a representant de l'interès públic, ja que el diàleg que es genera a les xarxes es pot equiparar a la conversa pública general. El volum es mostra a la figura 5.

Figura 5. Volum diari de piulades en anglès sobre el canvi climàtic



El pic correspon als dos dies en què es va celebrar la cimera. Font: http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Climate_Monitor_2015_0.pdf

A més, el caràcter textual dels missatges proporciona una opció que rarament es troba en altres mitjans: l'anàlisi de contingut i de sentiment. Tot i que l'estudi se centra en les piulades en anglès, castellà i francès, és important destacar que l'anàlisi automàtic de temes només es pot realitzar de manera adequada en el contingut en anglès. La ciència de dades avança molt ràpidament, però l'anàlisi del llenguatge natural requereix un esforç d'entrenament i de categorització manual que, avui dia, només és funcional per a l'anglès.

NLTK

La llibreria de Python més extesa i utilitzada per tractar el llenguatge natural és el Natural Language Toolkit (http://www.nltk.org/), però moltes de les seves funcions només són útils per a l'anglès.

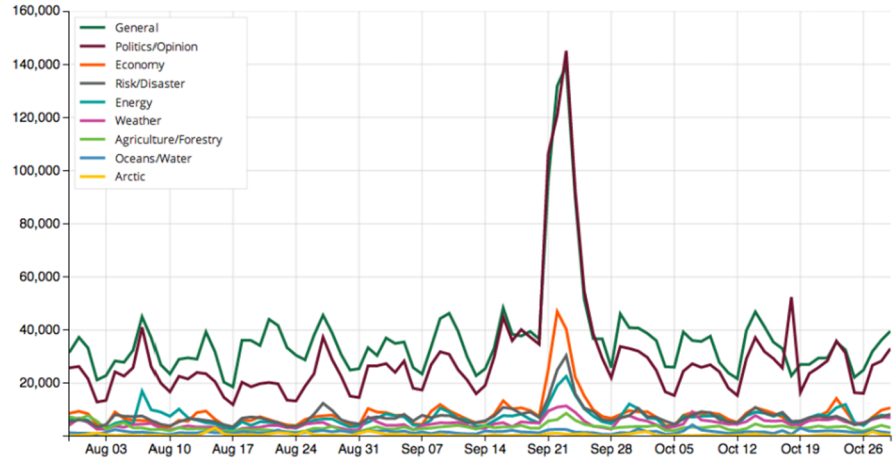
Els investigadors d'aquest cas van classificar els missatges en nou categories o temàtiques:

- general
- política/opinió
- energia
- economia

- riscos/catàstrofes
- agricultura/boscos
- temps (meteorològic)
- àrtic
- oceans/aigua

Cada missatge podia classificar-se en més d'un tema, si hi contenia referències. Per exemple, «la lluita contra el canvi climàtic comença protegint els oceans i els boscos» podria pertànyer fins a tres categories (general, oceans i boscos). El resultat gràfic de l'anàlisi es mostra a la figura 6.

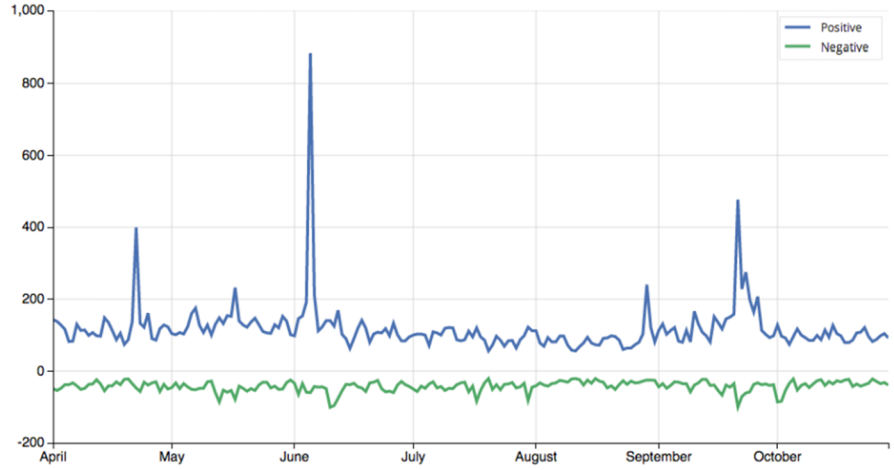
Figura 6. Volum diari de piulades en anglès segons el tema



Font: http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Climate_Monitor_2015_0.pdf

Les tècniques d'anàlisi de llenguatge natural no només ens permeten classificar temàtiques, sinó que també possibiliten, per exemple, deduir el sentiment associat als missatges. No és una ciència exacta ni fàcil: les persones escriuen amb estils diferents, amb sentits clars o irònics, amb elisions, amb frases complexes o simples. Així, l'anàlisi del sentiment és un intent estadístic de quantificar-ho. A la figura 7 es mostra l'anàlisi del sentiment. Per una banda, en positiu, els que parlen a favor d'actuar contra el canvi climàtic i per l'altra, els missatges que s'hi oposen.

Figura 7. Piulades positives i negatives sobre el canvi climàtic



Font: http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Climate_Monitor_2015_0.pdf

Combinant ambdós gràfics es poden extreure conclusions o percepcions durant el període d'estudi:

- De mitjana, cada dia s'escrivien uns 140.000 missatges relacionats amb el canvi climàtic. La gran excepció es produeix el mes de setembre, durant la celebració de la cimera, amb més de 400.000 missatges (multiplicant per tres la mitjana).

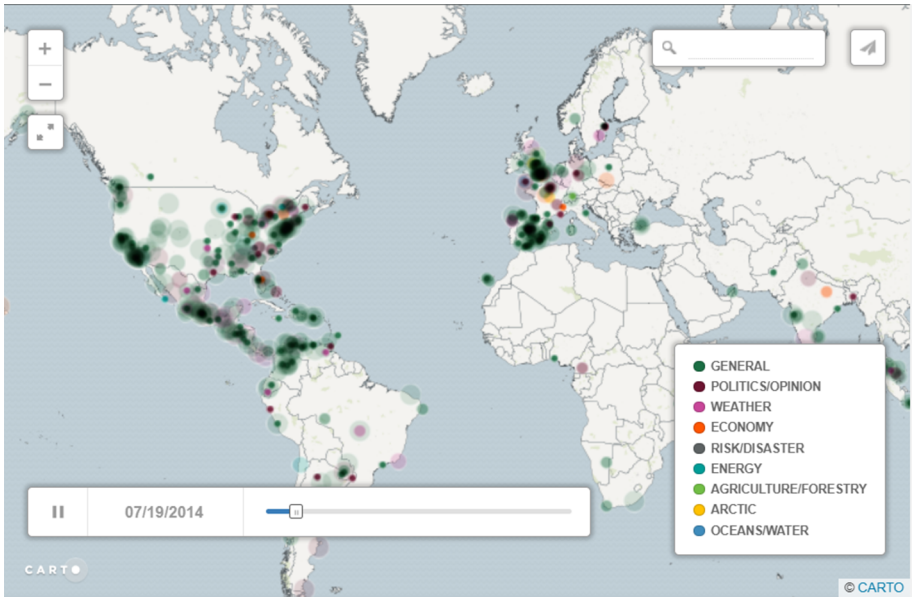
- La cimera va despertar interès durant l'esdeveniment, però també després. Durant el mes següent es van detectar entre el 10% i el 15% de converses addicionals.

- La política i l'economia són els temes que es comenten més. Els grans influenciadors de la conversa són, doncs, polítics i empresaris de renom. Aquestes observacions permeten, per exemple, decidir quines haurien de ser les personalitats amb qui s'hauria de contactar per aconseguir que el missatge tingués més impacte.

- El sentiment experimenta dos grans pics en els missatges polaritzats: el primer, el mes de juny, que coincideix amb la celebració del Dia del Medi Ambient, i el segon, el setembre, amb la cimera.

Així doncs, es comprova de manera metòdica com els esdeveniments (en aquest cas, una cimera) poden afectar el discurs i l'opinió pública. El monitoratge complet de la figura 8, amb mapa inclòs, es pot veure (i s'hi pot interactuar) a http://www.unglobalpulse.net/climate/.

Figura 8. Captura de l'animació dels missatges en el mapa



Font: http://www.unglobalpulse.net/climate/

2. Ciència de dades per al control d'epidèmies. El cas de l'Ebola

Les aplicacions de la ciència de dades per a salvar vides segueixen madurant i ho fan guiades tant pels avanços del sector com pels nous reptes que sorgeixen. Un d'ells és la gestió de les malalties infeccioses o epidèmies, en què la ciència de dades comença a oferir possibilitats tant a agències humanitàries com a ONG, ja sigui per veure tendències o correlacions, com per ajudar a la presa de decisions.

I és que la gran quantitat de dades derivades de l'ús de les tecnologies de la informació i la comunicació (TIC) mostra un potencial igualment elevat per enfrontar-se a aquests reptes. La petjada digital que deixa l'ús de serveis en línia, telèfons i altres transaccions digitals pot ser tractada, analitzada i utilitzada per millorar les decisions i per proveir serveis individualitzats amb informació personalitzada. En els països del Tercer Món, on les infraestructures poden presentar deficiències, l'expansió de l'ús del telèfon mòbil proporciona un valor particularment elevat, especialment en les emergències.

Un d'aquests projectes ha estat desenvolupat per la Unió Internacional de Telecomunicacions (ITU), l'agència de les Nacions Unides especialitzada en TIC. El treball de la ITU va consistir en utilitzar les dades massives per ajudar a seguir l'evolució d'una emergència sanitària respectant la privacitat dels usuaris.

ITU

L'objectiu de la Unió Internacional de Telecomunicacions és, literalment, «connectar tota la gent del món».

2.1. El cas de l'Ebola

L'any 2014 va ser l'any de l'expansió del virus de l'Ebola a l'est d'Àfrica, que va matar milers de persones. Els països més afectats van ser Libèria, Guinea i Sierra Leone, però el pànic i la por es van estendre a tot el món. Les agències sanitàries, a més, es trobaven amb la dificultat de traçar i contenir el virus mortal, especialment a causa del llarg període d'incubació de la malaltia i dels rituals funeraris d'alguns països.

La resposta de la ITU va ser el llançament de l'aplicació mòbil Ebola-Info-Sharing el 19 de desembre de 2014. Disponible en anglès i francès, aquesta aplicació gratuïta servia per distribuir informació oficial entre els usuaris i les organitzacions per facilitar la comunicació al terreny.

A més a més, la informació espaciotemporal que proporciona la població amb l'ús del telèfon mòbil és clau per a la intervenció en el control de l'Ebola. El motiu és simple: l'Ebola és una malaltia que es transmet per contacte i, per això, respon a la mobilitat humana. La globalització ha facilitat els desplaçaments i, per tant, també els contagis. Per altra banda, la tecnologia ha facilitat

les comunicacions. En conseqüència, les dades sobre les trucades (CDR, Call Detail Record) atreuen cada vegada més l'atenció de polítics i investigadors d'arreu per la seva capacitat de capturar els patrons de desplaçament humà.

Els CDR extrets de les xarxes mòbils possibiliten l'elaboració de mapes espacials que són un reflex de la vida diària de les persones d'un país o una zona, de les seves dinàmiques temporals i dels seus desplaçaments, i es converteixen així en una manera d'extreure o identificar problemes poc evidents. Una de les limitacions és, no obstant, el límit nacional de la majoria d'operadors, que fa difícil analitzar els moviments transnacionals.

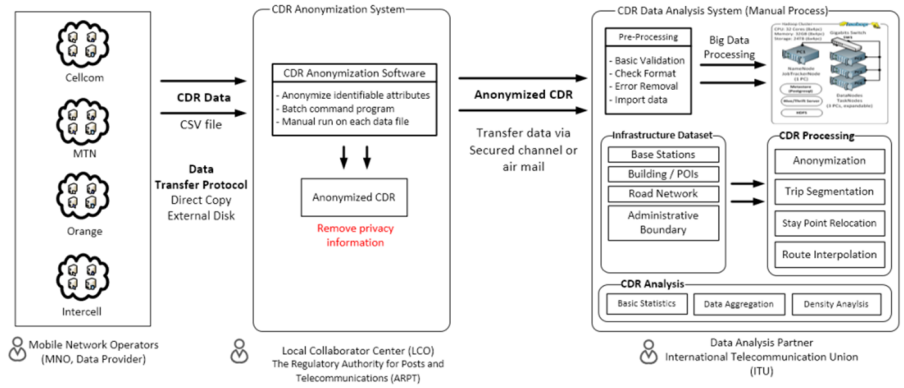
En aquest cas, els CDR s'apliquen per visualitzar i analitzar una informació difícil de tractar d'una altra manera: els moviments de les persones que s'han mogut des d'una zona afectada per un brot de la malaltia fins a altres zones. A partir d'aquesta informació es pot, doncs, preveure on és probable que hi hagi un nou brot, però també entendre les dinàmiques pròpies d'una població i contribuir a un millor pla d'acció per a reptes futurs.

2.2. Metodologia

Ens fixarem en el cas particular de Guinea[1]. El procés d'anàlisi dels CDR necessita múltiples passos que es mostren a continuació (figura 9):

(1)https://bit.ly/2PIG2bJ

Figura 9. Procés d'anàlisi de dades de la informació dels telèfons mòbils



Font: https://bit.ly/2PIG2bJ

- Operadors de telefonia mòbil: la informació es recull mitjançant els operadors. En aquest cas, la informació de posicionament s'obté de les dades que té l'operador de cada terminal en els seus servidors. Aquest conjunt de dades s'exporta als centres col·laboradors locals i normalment s'obtenen en format CSV (separat per comes). En aquest cas, les transferències es van fer de manera manual amb còpies en discs durs externs. No tot és alta tecnologia!

- Centres col·laboradors locals: un pas important, especialment per a l'ètica i la privacitat, és l'eliminació de les dades personals. Aquestes unitats s'encarreguen de fer la neteja i són habitualment liderades pel mateix re-

gulador, que s'encarrega de passar dades anonimitzades (i, per tant, no identificables) a la unitat que s'encarregarà de l'anàlisi de dades.

- Centre d'anàlisi de dades: el rol d'aquest centre o col·laborador és el d'emmagatzemar, mantenir i tractar les dades obtingudes. Les dades es passen a un entorn de dades massives (vegeu-ne l'arquitectura a la figura 10). L'anàlisi posterior consisteix a aconseguir:
  - Dividir els viatges i desplaçaments dels usuaris en segments, de manera que es puguin identificar els trajectes en moviment i els llocs d'aturada.
  - Buscar els punts d'estada. Com que les dades obtingudes proporcionen la posició de la torre més propera no són prou exactes per situar l'usuari, però si es treballa amb les localitzacions en una zona concreta és possible determinar la posició més probable del subjecte a partir de la triangulació. Amb les dades de diverses torres, en canvi, és possible trobar més punts de referència i, per tant, limitar la zona on és possible que es trobi l'usuari. Així s'aconsegueix una localització més realista.
  - Interpolar les rutes, ja que una vegada determinats els punts d'estada s'ha de dibuixar la ruta més probable entre les dues, que no sempre coincideix amb el camí més curt.
  - Agregar els resultats en una graella. En aquest cas, per hora i per quilòmetre quadrat (no es necessita més granularitat).
  - Visualitzar les dades. En aquest cas, es va fer un mapa animat utilitzant Mobmap[2]

csv

El format CSV (fitxers separats per comes) és un format habitual d'organització de grans quantitats de dades, especialment en empreses, encara que comença a estendre's l'ús dels JSON (JavaScript Object Notation), molt més flexible i llegible.

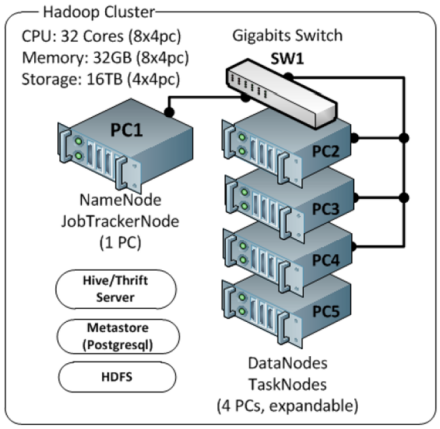
[(2)]Podeu provar de carregar unes dades de mostra sobre Tokyo a:

http://shiba.iis.u-tokyo.ac.jp/member/ueyama/mm/app/?load=opf.

2.3. Resultats

Com a resultat de l'estudi es van extreure moltes dades i conclusions interessants, des d'una aproximació de cens de telèfons mòbils al país fins a una correlació entre usuaris, zones més habitades i desplaçaments més habituals. També es va veure, per exemple, quins eren els recorreguts que se solien fer entre pobles. Però per al cas que ens ocupa ens centrarem en una part més interessant: el dibuix de l'Ebola.

Figura 10. L'arquitectura de les dades massives utilitzada. Hadoop no és ciència-ficció!



Font: https://www.itu.int/en/ITU-D/Emergency-Telecommunications/Documents/2017/Reports/GN/EN/D012A0000D03301PDFE.pdf

Mitjançant l'anàlisi i la visualització és possible fer coses com les de les figures 11 i 12. En un primer cop d'ull es veu una zona afectada per l'epidèmia i que, per tant, es considera zona de perill. Els punts vermells indiquen persones que es trobaven dins de la zona en el moment de perill, mentre que els punts petits i blancs són altres persones que es consideren fora de perill.

Figura 11. Persones que es troben en una zona de risc

Font: https://www.itu.int/en/ITU-D/Emergency-Telecommunications/Documents/2017/Reports/GN/EN/D012A0000D03301PDFE.pdf

Amb aquesta informació (i marcant els punts que tenen una probabilitat més alta d'haver entrat en contacte amb la malaltia) es pot traçar l'evolució de l'Ebola al llarg del temps. La figura 12, per exemple, mostra l'evolució i els camins fets per aquests punts 48 hores després d'haver passat per la zona de perill. Fixeu-vos que lluny han arribat alguns dels punts. En particular n'hi ha tres que ja són als extrems del país, amenaçant una àrea enorme de Guinea en el seu recorregut.

Figura 12. Desplaçaments de les persones analitzades 48 hores després de passar per la zona de perill

Font: https://www.itu.int/en/ITU-D/Emergency-Telecommunications/Documents/2017/Reports/GN/EN/D012A0000D03301PDFE.pdf

Si bé la falta d'informació immediata i exacta sobre els moviments de les persones en una catòstrofe natural, emergència o epidèmia, pot limitar seriosament l'efectivitat de la resposta humanitària, hem vist com la naturalesa ubiqua dels telèfons mòbils es pot convertir en una oportunitat. L'activitat de trucades, missatges i l'ús de les torres de comunicació en general proporciona informació valuosa de l'activitat de les persones (acumulacions i desplaçaments després o durant un esdeveniment), per així millorar els avisos a la població i la gestió d'emergències. Casos com l'Ebola, a més, són epidèmies difícils de controlar, fàcilment contagiables i que els seus portadors poden estendre ràpidament a escala nacional o internacional, així que cal tenir maneres per actuar ràpidament. Com acabem de veure, la ciència de dades hi pot tenir un paper destacat.

3. L'anàlisi i la visualització dels recorreguts dels taxis de Nova York

Fins ara hem vist casos reals d'aplicació de la ciència de dades però no hem pogut seguir el seu procés complet. És el moment de veure un exemple complet que, a més, es basa en un conjunt de dades accessible i del dia a dia: els recorreguts que fan els taxis a la ciutat més famosa del món, Nova York.

3.1. Dades

Cada dia que passa s'acumulen més i més dades a les organitzacions; la majoria queden en mans privades (la qual cosa és normal, ja que la informació es pot considerar un avantatge competitiu), però també hi ha conjunts de dades que es posen a disposició del públic general. Hi ha fins i tot organitzacions que opten pel concepte de dades obertes i permeten l'ús i aprofitament de les seves dades per part de qualsevol persona interessada.

En el cas dels viatges amb taxi que ens ocupen, no es tracta exactament de dades obertes; una llei de l'estat de Nova York, no obstant, permet demanar dades sota certs supòsits. Amb una petició d'aquestes característiques es va obtenir un conjunt de dades molt detallat que conté tots els viatges realitzats per taxis l'any 2013[3]. Entre les dades s'inclou el punt de recollida (inici del viatge) i el punt final (o destinació), el temps del trajecte, la distància recorreguda i el cost.

(3)Podeu llegir la història sencera a la pàgina web del seu autor: https://chriswhong.com/open-data/foil_nyc_taxi/.

Notebook

Podeu seguir l'explicació en paral·lel al Notebook a: https://bit.ly/2IxZKnP.

El fet de ser dades reals (i grans) es fa palès, sobretot, en dos aspectes. El primer, que ocupa més de 19 GB repartits en diversos fitxers separats per comes[4] i amb 14 milions de registres per fitxer. El segon, que hi ha una gran quantitat de registres incomplets, errors, columnes supèrflues, etc.

(4)De fet, les dades estan disponibles aquí: www.andresmh.com/nyctaxitrips/.

En aquest mòdul seguirem un procés típic dels projectes de ciència de dades: analitzarem la informació, construirem un model que avaluarem i intentarem fer prediccions. I diem que ho intentarem perquè, qui sap, potser aquest cas amaga un secret més fosc del que sembla...

3.2. L'anàlisi de la informació

En primer lloc, cal saber de quines dades disposem. Per donar una primera ullada, el més senzill és carregar les dades i veure-les en forma de taula. En el cas dels taxis, per exemple, les primeres files i columnes tenen la forma que es mostra a la taula 1.

Taula 1

	medallion	*hack-license*	*vendor-id*	*rate-code*	*...*	*trip_dist*	*pickup_lat*	*...*
0	89D227...	BA96D...	CMT	1	...	1.0	40.757977	...
1	0BD7C8...	9FD8F...	CMT	1	...	1.5	40.731781	...
2	0BD7C8...	9FD8F...	CMT	1	...	1.1	40.737770	...
3	DFD220...	51EE8...	CMT	1	...	0.7	40.759945	...
4	DFD220...	51EE8...	CMT	1	...	2.1	40.748528	...

Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Aquesta és només una mostra per fer-se'n una idea, ja que la taula és molt més gran (i extensa). Anem per parts:

- Les dues primeres columnes (*medallion* i *hack-license*) són columnes identificadores que identifiquen la llicència del taxi i que, per tant, no són gaire interessants a l'hora de construir el model.

- Més enllà trobem columnes com *vendor-id* o *rate-code*, que semblen ser variables categòriques. Sol ser interessant representar aquestes variables gràficament per entendre'n les distribucions, com veurem a continuació. En aquest conjunt de dades hi ha variables per identificar a quin grup empresarial pertany el taxi, quin codi de tarifa es va aplicar o quin mitjà de pagament es va utilitzar, per exemple.

- També tenim una sèrie de columnes amb uns valors molt més familiars: nombres. Aquestes columnes tenen dades sobre la distància recorreguda en el viatge, el cost o la durada. Amb les variables numèriques també és interessant veure si es correlacionen entre elles; els gràfics més habituals en aquest cas són els núvols de punts o *scatterplots*.

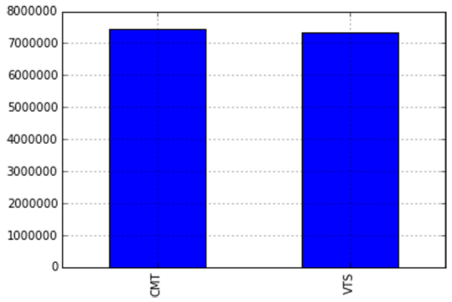
- Finalment, el conjunt de dades té una sèrie de xifres que, tot i ser numèriques, corresponen a un domini molt concret: les coordenades geogràfiques dels punts de sortida i d'arribada, en latitud i longitud. La particularitat d'aquestes columnes és que els seus valors es limiten a la geografia estudiada però, sobretot, que es poden representar en un mapa, aportant informació addicional.

A continuació ens centrarem en algunes de les variables per entendre el procés inicial.

3.2.1. Variables categòriques

Una de les primeres coses que es poden fer és visualitzar la distribució de les variables categòriques, per intentar veure'n la rellevància. La primera, per exemple, és el proveïdor del sistema (figura 13).

Figura 13. Representació de la freqüència de cada *vendor*

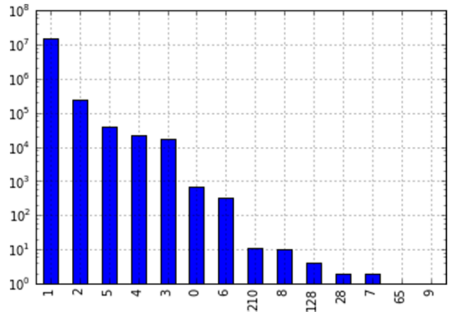


Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Com es pot observar, el conjunt de dades analitzat és només una part del total, que conté una mica més de 14 milions de registres. D'aquests, més o menys la meitat pertanyen a cada un dels serveis (CMT i VTS).

És més interessant veure les zones tarifàries. Del gràfic de barres de la figura es pot extreure que la zona 1 és desproporcionadament present. De fet, cal fixar-se en que, per a una correcta representació, l'escala de l'eix vertical (que representa la freqüència absoluta d'aparició de cada tarifa) s'ha hagut de fer logarítmica. Així, una primera deducció és que, tenint en compte que és Nova York, és més que probable que la zona 1 correspongui a l'illa de Manhattan.

Figura 14. Zones tarifàries dels viatges del conjunt de dades

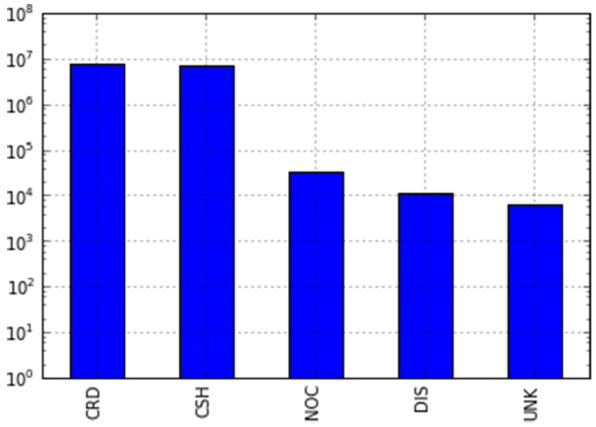


Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Una altra observació interessant correspon al mètode de pagament, representat a la figura 15. I és interessant no tant perquè veiem que els dos mètodes més habituals, amb diferència, son la targeta de crèdit (CRD) i l'efectiu (CSH), sinó perquè podem detectar que, per exemple, hi ha certa quantitat de dades desconegudes (UNK). En aquest cas es tracten d'uns pocs milers de línies que,

en una mostra de 14 milions, poden ser poc significatives, però serveix com a exemple de dades que ens hauríem de plantejar si s'haurien de netejar abans de seguir elaborant un model.

Figura 15. Mètodes de pagament

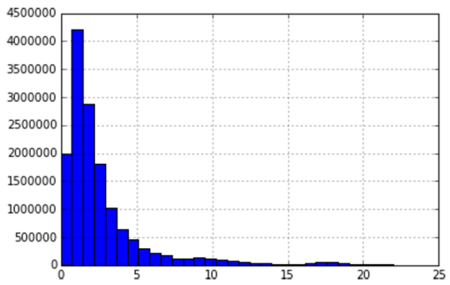


Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

3.2.2. Variables numèriques

Per a les variables numèriques ja no farem gràfics de barres, sinó que mostrarem la seva versió contínua: histogrames. Si tenim en compte que estem analitzant viatges amb taxi, què pot ser interessant? Doncs la distribució de les distàncies recorregudes, per exemple (figura 16).

Figura 16. Histograma de les distàncies recorregudes

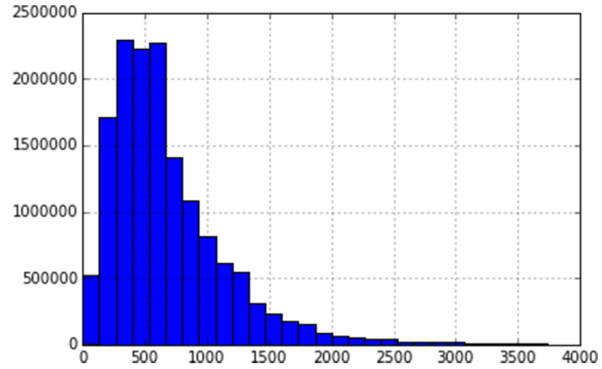


Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Així, comprovem com la distribució tendeix cap a viatges curts però fixeu-vos que els trajectes que una persona pot fer a peu es fan poc amb taxi. La majoria de trajectes són entorn d'un quilòmetre de distància, tot i que n'hi ha de molt més llargs.

Seguint la mateixa línia podem comprovar els temps de trajecte. A la figura 17 podem apreciar la distribució, que mostra com els trajectes més habituals se situen entorn dels 500 a 700 segons, més o menys 10 minuts de viatge.

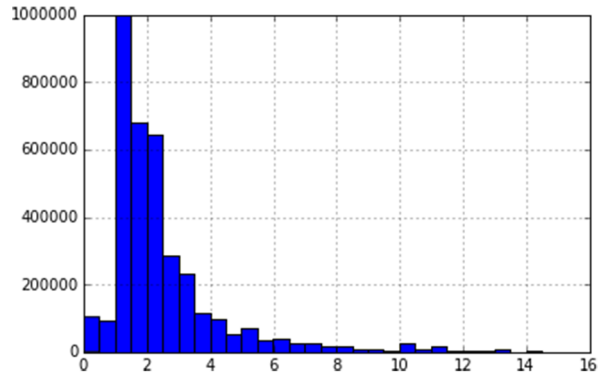
Figura 17. Histograma de la durada dels viatges



Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Fins ara hem vist variables molt descriptives (i no deixen de ser interessants), però a continuació ens fixarem en una que, com a taxistes, potser ens resulti més important: les propines. Imaginem per un moment que el nostre objectiu final és entendre què motiva un client a deixar propina i de quina quantitat. Podem començar per visualitzar les propines en els viatges en què se'n dona (que no són tots).

Figura 18. Distribució de les propines



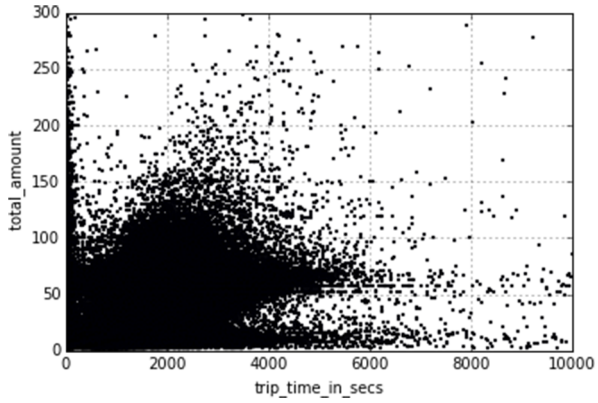
Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

De la figura 18 s'extreu que la propina més habitual està entre un i dos dòlars (per la qual cosa intuïm que aquest fet també està relacionat amb el fet que la majoria de viatges siguin curts). No és gaire informació, però com a mínim veiem que cada dòlar compta i que, per tant, cal ser eficient en els viatges.

Abans hem indicat que les variables quantitatives permeten veure les seves relacions. A la figura 19, per exemple, es mostra el núvol de punts de la relació entre el temps de viatge i el cost. Fixeu-vos en diverses coses importants. En primer lloc, es veu certa correlació (que s'espera) entre el temps i el cost (la idea és que com més temps, més car és un viatge). No obstant, la dispersió es bastant gran i, per tant, la relació no és clara. Però un gràfic com aquest també ens permet veure altres coses. Us heu fixat que sobre l'eix vertical hi ha molts punts? Sembla ser que el nostre conjunt de dades té molts viatges d'una durada de 0 segons. No té gaire sentit, oi? Doncs encara en té menys quan

alguns d'aquests viatges de poca durada tenen, a més, costos molt alts. És un clar indicador de problemes en alguns registres, que poden contenir errors i necessitar ser eliminats del conjunt abans de seguir.

Figura 19. Relació entre la propina i el temps de viatge

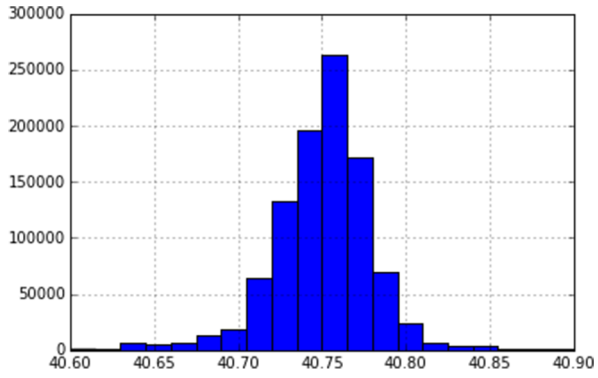


Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

3.2.3. Variables geogràfiques

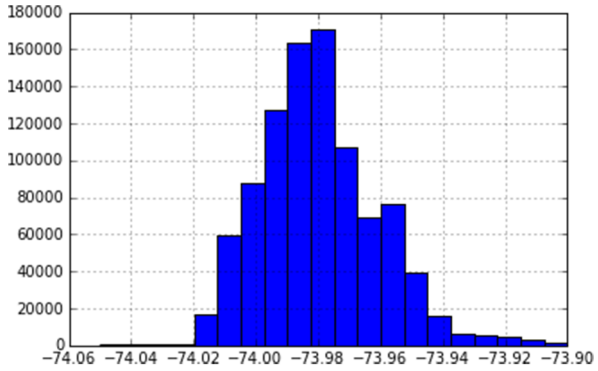
Durant les observacions inicials també hem detectat tot un seguit de variables geogràfiques. Aquestes incloïen la latitud i la longitud, tant de recollida com de destinació. Si representem l'histograma de les latituds i longituds dels punts de destinació, per exemple, obtindríem els histogrames de les figures 20 i 21, que mostren una distribució similar a una de normal.

Figura 20. Latitud dels punts de destinació



Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

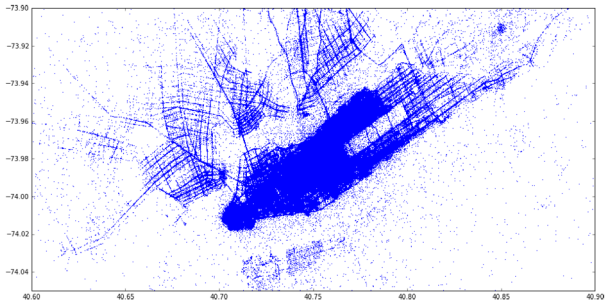
Figura 21. Longitud dels punts de destinació



Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

No obstant, les coordenades geogràfiques tenen més informació que un simple histograma. Les podem dibuixar o situar en un mapa. O en un cas amb tantes dades com aquest, utilitzar-les perquè siguin els mateixos punts els que facin un patró i que construeixin el mapa. Fixeu-vos en quin gràfic més bonic (i informatiu) s'obté de representar les parelles longitud-latitud dels punts de recollida (figura 22).

Figura 22. Mapa generat a partir dels punts de recollida



Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

No hem obtingut només un mapa de Nova York, sinó que a més són fàcils de detectar les zones amb més activitat, marcades amb un color més intens i dens. El perfil de Manhattan s'intueix a la perfecció i el color es va difuminant a l'allunyar-se del centre de la ciutat (és a dir, s'agafen més taxis al centre que a la perifèria).

3.3. La creació d'un model

Imaginem ara que som taxistes. Després de visualitzar i analitzar el conjunt de dades ens hem adonat d'una columna que ens desperta l'interès: la propina. Durant la nostra feina diària hem vist que alguns clients no deixen propina i altres sí que ho fan, que de vegades deixen una quantitat petita i altres, una de més substancial. Més d'una vegada hem donat voltes als factors que fan que un usuari deixi una quantitat determinada de propina, però mai no hem tingut prou dades per intentar-ho entendre. Mai... fins ara.

Una vegada arribats a aquest punt no cal que frenem l'ambició. Si fóssim capaços de conèixer els factors d'influència potser també seríem capaços de preveure quins clients són més propensos a deixar bones propines i, per tant, podríem intentar recollir sempre els millors clients. Fins i tot, podríem instal·larnos una aplicació al telèfon mòbil que ens avisés dels clients garrepes abans que fos massa tard i ja haguessin pujat al cotxe.

Així, els objectius podrien ser:

- Entendre els factors que influeixen sobre deixar o no propina en els viatges amb taxi a Nova York.

- Utilitzar aquest coneixement per predir la propina que rebrem i, per tant, poder evitar les situacions sense propina.

Aquest cas, no obstant, inclou una lliçó important. Si intentéssiu construir un model amb el conjunt de dades complet us sortirien uns resultats molt bons en les mètriques, però quan intentéssiu aplicar-lo al vostre taxi els resultats serien un desastre. El motiu és que les prediccions massa bones per ser veritat no existeixen.

Una regla no escrita de la ciència de dades és que si s'obté una precisió més gran que l'esperada, el més probable és que el model estigui fent alguna cosa inesperada. El món és complex i difícil, seria estrany que es pogués modelar de manera assequible. En aquest cas, el model inicial apareix amb una variable dominant en la predicció de la propina: la forma de pagament.

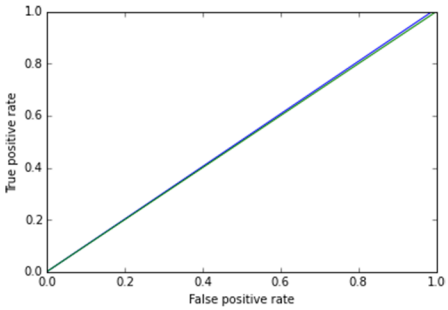
Com a usuaris ocasionals del taxi podríem pensar que el normal són les situacions següents: els clients que paguen amb targeta és més probable que utilitzin la moneda electrònica per pagar el cost exacte i no deixar propina; les persones que paguen en metàl·lic, en canvi, solen arrodonir el resultat del taxímetre i, per tant, sempre deixen alguna cosa de propina. Té sentit, oi? Doncs no.

Si agafem el conjunt de dades i fem les proves, veurem com la immensa majoria dels usuaris que paguen amb targeta deixen propina. I què passa amb els usuaris que paguen en metàl·lic? El nostre conjunt de dades diu que... ningú ha deixat propina! Com pot ser això? Doncs, simplement, no és possible. L'explicació és senzilla: quan els clients deixen propina en metàl·lic, el conductor no ho registra de la manera necessària perquè aparegui a les dades. Tantes ganes de saber si un client deixa o no propina i al final sembla ser que el que acabem de descobrir és un clar cas de frau a la hisenda pública!

Deixant les curiositats a banda, quan ens trobem en un cas així no hi ha mitges tintes: si les dades estan compromeses, no es poden utilitzar. Així, podem refer la pregunta i intentar buscar els motius que porten a clients **que no paguen en efectiu** a deixar propina. Ens sap greu eliminar la meitat dels registres, sí, però no tenim altre remei.

Amb les dades netes ja podem passar a construir el model. Una manera fàcil de començar és amb un model de regressió logística per intentar classificar els viatges segons la propina. No entrarem en el detall de la implementació (podeu consultar-la al Notebook), però sí que veurem els resultats de la corba ROC (figura 23). Aquesta corba representa la precisió del model i, amb un valor d'aproximadament 0.5, indica que el model no és millor que una predicció aleatòria del resultat. És a dir, que si triéssim els viatges a sorts encertaríem igual que el model. Però no ens desanimem!

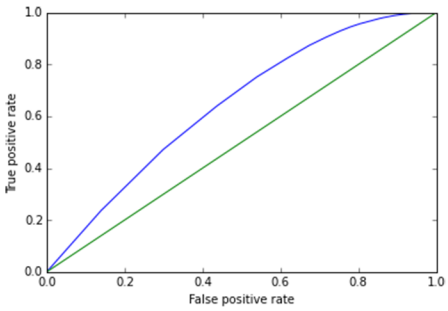
Figura 23. Corba ROC del classificador lineal



Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Veient els mals resultats del classificador lineal, podem passar a altres opcions. La més directa és buscar un classificador que no sigui lineal. En aquest cas s'utilitza un model de bosc aleatori (*random forest*), que com hem vist en mòduls anteriors, no és més que un conjunt d'arbres de decisió.

Figura 24. Corba ROC del classificador no lineal



Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

Aquest model obté la corba de la figura 24. El més important a tenir en compte ara mateix és que l'àrea que queda sota la corba és del 0.64, millorant notablement el rendiment del model anterior. Encara faltaria molt camí, però per ara ho deixarem aquí. Aquest model ja ens pot ser útil, ja que ens indica que hi

ha certa tendència —és a dir, que hi ha algunes variables que tenen influència sobre la propina— i ens pot ajudar a identificar-les. I és que segons aquest model, les variables més importants són les que es mostren en la taula.

Taula 2

	Característica	Importància relativa
0	dropoff_latitude	0.165411
1	dropoff_longitude	0.163337
2	pickup_latitude	0.163068
3	pickup_longitude	0.160285
4	trip_time_in_secs	0.122214
5	trip_distance	0.112020
6	fare_amount	0.067795

Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

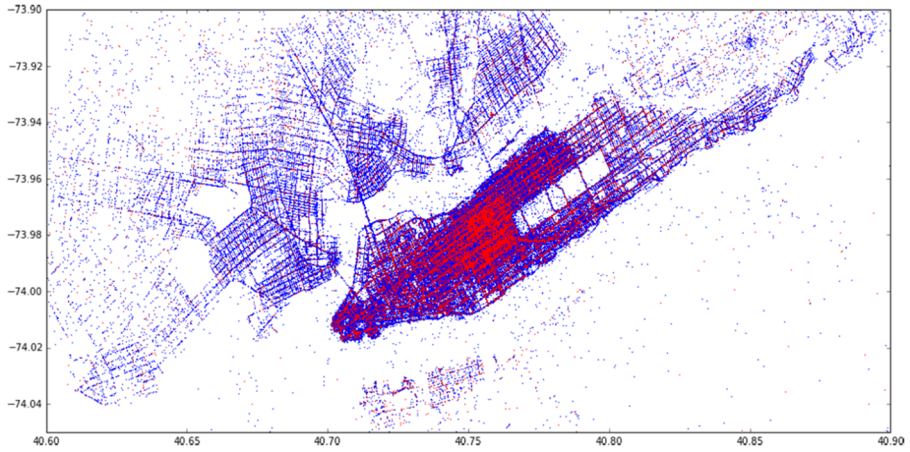
Fixeu-vos que sembla que l'origen i la destinació de l'usuari són la part més rellevant, seguits de tres variables que podrien resultar més obvies: la durada, la distància i el cost del viatge. Fins ara, però, no hem introduït variables categòriques al model; aquest seria el moment de començar a introduir factors addicionals. Podríem, per exemple, introduir una diferenciació en funció de la zona tarifària o del mètode de pagament (sempre recordant que hem deixat fora el pagament en metàl·lic). I això no és tot, en ciència de dades també és molt habitual fer enginyeria de característiques o *feature engineering*. Aquesta enginyeria consisteix a generar noves característiques a partir de les ja existents. Podríem crear, per exemple, una mesura de velocitat mitjana (dividint la distància pel temps) o aprofitar les variables de data i temps per extreure el dia de la setmana (o si és dia laborable o festiu) o la hora (per si és hora punta). Les possibilitats són infinites però sempre cal fer-ho amb seny, les noves variables han de tenir sentit i hem d'intentar que no estiguin relacionades directament amb les variables del conjunt. En cas contrari, sigui per relació o per excés, podem tenir problemes de sobreentrenament amb facilitat.

3.4. L'obtenció de coneixement

Tot i que l'obtenció del model hagi resultat un fracàs relatiu, el simple acte de construir-lo ens ha aportat coneixement. De fet, la conclusió més important encara no l'hem condensat d'una manera visible. A la llista de característiques rellevants de la taula 2 han aparegut les coordenades com a les variables més importants. És a dir, sembla ser que les propines depenen, sobretot, de la seva distribució geogràfica. Potser aquí tenim una pista important si som taxistes; sembla que hi ha llocs on, si recollim un client, és més fàcil que ens deixin propina. Quins seran? Times Square? A continuació ho descobrirem.

La figura 25 mostra els punts de destinació amb la peculiaritat que pintem de vermell els punts dels viatgers que no deixen propina i de blau els que sí que en deixen. Hi veieu alguna cosa interessant?

Figura 25. Distribució geogràfica de les destinacions



De color vermell, els que no deixen propina i de color blau, els que sí que en deixen. Font: https://github.com/brinkar/real-world-machine-learning/blob/master/Chapter%206%20-%20NYC%20Taxi%20Full%20Example.ipynb

La conclusió és clara: els viatgers que acaben el seu trajecte al centre de la ciutat no solen deixar propina. I per què ocorre això? Hi ha diverses possibilitats:

- Pot ser que la congestió del trànsit faci que els passatgers es desesperin perquè arriben tard, els conductors facin sonar el clàxon i s'estressin i, en general, que el viatge sigui menys agradable.

- Però també pot ser que sigui culpa dels turistes. La majoria de taxis del centre els utilitzen visitants estrangers que, en gran mesura, venen de països europeus (que rarament deixen propina) o asiàtics (que encara en deixen menys). No és que siguin més garrepes, sinó que és una qüestió cultural.

Així, un taxista que vulgui més propines no ha de fer res més que anar als afores a recollir passatgers. El problema serà que, probablement, no reculli tants viatgers, però bé aquí buscàvem informació sobre la propina, no sobre el benefici net. Al final allò important és veure com les dades generades en el món real ens poden servir per explicar fenòmens sobre el mateix món real i les persones que generen aquesta informació.

4. Resum

En aquest mòdul hem vist diversos casos d'aplicació de la ciència de dades, casos que extreiem del món real i que intenten buscar resposta a problemes rellevants. Hem repassat també quin és el procés d'un projecte complet de ciència de dades, tot i que no s'hagi entrat en el detall del codi. El més important és que quedin clars els punts següents:

- Cada dia hi ha més organitzacions i individus produint quantitats de dades cada vegada més grans.

- Alguns d'aquests conjunts de dades estan disponibles públicament, d'altres són privats. Els seus objectius són diversos, des de millorar la vida de les persones a obtenir benefici, passant per obtenir coneixement sobre alguns fenòmens i altres casos.

- Les dades del món real solen ser complexes, poc netes i incompletes. Visualitzar-les ajuda, de la mateixa manera que ajuda conèixer el context de les dades.

- Els resultats que són massa bons per ser certs probablement no ho siguin.

- Normalment es comença per models simples als quals es va afegint complexitat per millorar-los. És important tenir clar el motiu que hi ha darrere de cada decisió.

- No hi ha cap pas inútil, fins i tot un model fracassat pot servir per obtenir coneixement.

Bibliografia

Brink, Henrik; Richards, Joseph W.; Fetherolf, Mark (2016). *Real-World Machine Learning*. Nova York: Manning Publications Co.

UN Global Pulse (2017). *Inferring Jakarta Commuting Statistics from Twitter*. <https://www.unglobalpulse.org/inferring-jakarta-commuting-statistics-twitter>.

UN Global Pulse (2017). *Using Big Data to Study Rescue Patterns in the Mediterranean*. https://www.unglobalpulse.org/projects/using-big-data-study-rescue-patterns-mediterranean>.

UN Global Pulse (2015). *Using Twitter to Measure Global Engagement on Climate*. <https://www.unglobalpulse.org/projects/Twitter-Climate-Change>.

Zavazava, Cosmas (2015). *How Big Data will help fight global epidemics*. <https://news.itu.int/big-data-will-help-fight-global-epidemics/>.