Fonis Datageeks

Homework: Applied Machine Learning

Zdravo datageekovi,

Ovaj domaći predstavlja kratki test svega što smo do sada radili i prvi kompletan data science projekat. Na našem drajvu nalazi se folder sa podacima.
Potrebno je izabrati jedan dataset od ponuđenih i nad njim sprovesti kompletnu analizu. Svi datasetovi su preuzeti sa interneta, pa s toga budite slobodni da guglate ako su vam potrebne dodatne informacije za razumevanje podataka.
Ovaj domaći jeste kompletan data science projekat, budite slobodni da eksperimentišete, probate sve što smo radili ili nismo. Ispod slede smernice i pitanja, koja nisu konačna, pa zato razmislite o tome šta biste dodatno mogli bolje da uradite.

1. Eksploratorna analiza podataka

Nad izabranim datasetom potrebno je izvesti **eksploratornu analizu**, objasniti prirodu podataka i glavne pravilnosti uočene tokom analize. Kombinovati statističke mere i vizualizaciju kako bi objašnjenja bila što preciznija. Nije potrebno objašnjavati mere i grafike, već rezultate. Obratiti pažnju da rečenice budu koncizne i da se odnose na zaključke otkrivene tokom analize. Zaključci ne bi trebalo da sadrže nešto što vlasnik podataka već zna.

2. Priprema podataka

1. Proveriti da li postoje nedostajuće vrednosti. U slučaju da postoje, odlučiti se za strategiju rada sa njima (možete izbaciti te redove, popuniti ih srednjom vrednošću, minimumom, maksimumom (...) u zavisnosti od toga šta zaključite da znače te nedostajuće vrednosti). Dobra referenca.
2. Pronaći izuzetke (outliers) u podacima. Prikazati dataset izuzetaka i kratko objasniti koji je izvor njih (stvarni izuzeci (nepredviđene situacije), greške u zapisu podataka…). Ukloniti izuzetke iz dataseta.
3. Obratiti pažnju na kategoričke osobine. Koristeći različite metode, pripremiti ih za model. Razmisliti o tome koje osobine su ordinalne a koje nominalne, za osobine sa velikim brojem mogućih vrednosti razmisliti o kreiranju nove kategoričke osobine čiji bi skup vrednosti bio manji.
4. Skalirati numeričke osobine. Probati bar dva različita načina za skaliranje numeričkih osobina i odlučiti se za jedan (okej je skalirati neke atribute na jedan, a druge na drugi način). Objasniti kako funkcioniše izabrani metod za skaliranje atributa.

5. Opciono, izvršiti selekciju bitnih osobina: izbaciti visoko korelisane osobine ili one koje ne utiču na target (moguće uraditi i kasnije).
6. Da li u podacima postoji nebalansiranost? Izabrati jednu tehniku za rešavanje nebalansiranosti, primeniti je i objasniti. Veliki plus je ako se koristi metoda koja nije rađena na radionici.

3. Izgradnja prediktivnog modela klasifikacije

1. U jednoj rečenici, definisati problem koji se rešava i meru evaluacije koja se prati.
2. Probati bar 3 različita algoritma klasifikacije (veliki plus je ako se koristi neki algoritam koji nije rađen na radionicama, još veći ako se on objasni u radu). Koristiti kros-validaciju za optimizaciju parametara. *Podrazumeva se, pazite na under/overfitovanje. ;)*
3. Odlučite se za jedan algoritam, njegove parametre i izgraditi konačan model. Objasniti izabrani algoritam tako da "neko ko nije u oblasti" može da razume intuiciju iza njega.

4. Evaluacija modela

Tokom prethodnog koraka neophodno je koristiti neku od mere evaluacije. U ovom koraku potrebno je izmeriti više mera evaluacije nad konačnim modelom. Na osnovu razumevanja poslovnog problema, odlučiti si za najbitniju i objasniti odluku.
Veliki plus je ako se koristi neka od mera koja nije rađena na radionicama. U tom slučaju, potrebno je objasniti je. Moguća referenca i uz to sklearn dokumentacija.

Još par informacija

Ne sviđaju ti se datasetovi? Nađi bolji na Kaggle-u ili UCI ML Repository, pošalji mi i biće dodat ako je odgovarajući.

Obratiti pažnju na kom skupu podataka se radi svaki od navedenih koraka (trening, test, izuzeci, kros-validacija itd.).

Poslati sređeni notebook, različite delove jasno odvojiti naslovima (razumno korišćenje h1,h2, h3..).

Važno je da rad ima čitav proces, odnosno sva 4 navedena koraka, ali je okej ako ne uspeš da ispoštuješ sve navedene zahteve i uradiš na malo jednostavniji način. Maksimalno jednostavan rad bi bio poput onog prikazanog na 3. radionici u notebooku Klasifikacija.

Svaki rad biće ocenjen:
1. od strane drugih polaznika radionica,
2. sa strane ispunjavanja svih osnovnih i dodatnih zahteva i
3. sa strane postignute tačnosti.

Ovo je pravi mini projekat i zahteva energiju, vreme i posvećenost. Preporuka je da se čujete sa drugarima, razmenjujete zaključke i postavljate pitanja jedni drugima što više. Skroz je okej i da zajedno radite i pomažete jedni drugima, ali svako mora imati svoj projekat. Takođe, veoma je poželjno da u grupi postavljate pitanja u vezi nedoumica koje imate i odgovarate na tuđa pitanja.

Kada budemo imali dovoljan broj urađen radova, možemo organizovati druženje gde bi svako predstavio svoje rezultate.

Pokidajte! :)