# Grupiranje dvije osobe unutar istog kućanstva na temelju njihove međusobne interakcije

## Ulaz u model

Vektor značajki predstavlja svu zabilježenu interakciju između korisnika A i korisnika B. Vrlo se lako generira iz dostupnih podatka o korisnicima I njihovoj interakciji.

## Značajke modela

Parametri distribucije trenutka kada se dogodila interakcija između korisnika A i B na razini dana u tjednu.

(npr. ako postoji 100 poziva i 20 SMS-a između korisnika, računa se distribucija vremena kad su se dogodile te interakcije. Distribucija mora biti određena unaprijed, te se ona može zaključiti iz neoznačenih podataka. Prilikom generiranja vlastitog skupa podataka smo pretpostavili normalnu distribuciju jer smo uočili da je to najčešća distribucija u sličnim javno dostupnim skupovima podataka).

Parametri distribucije duljine razgovora između dvije osobe na razini dana.

IOU (engl. Intersection over union) (gdje se gleda i frekvencija) lokacija osoba. Ova značajka predstavlja postotak zajedničkih lokacija kroz ukupan broj lokacija gdje su se dogodile interakcije između korisnika, ali samo onda kada trenutna lokacija između korisnika nije ista. Ova značajka je osmišljena s pretpostavkom: ako osoba A i osoba B žive na istoj lokaciji postajati će preklapanja u lokaciji osoba (osoba A doma zove osobu B na poslu i obrnuto). Ali prije računanja se filtriraju interakcije s iste lokacije tako da se izbjegnu situacije kada dvoje kolege unutar firme razgovaraju jer su na različitim katovima i slične situacije. Ova značajka omogućava detekciju ljudi koji su u VIP mreži ali koriste prepaid bonove.

Binarna značajka koja govori jesu li obje osobe u VIP-u ili je jedna od osoba izvan VIP-a.

## Izlaz modela

Klasifikacija pripadaju li osoba A i osoba B istom kućanstvu.

## Isprobani modeli

U sklopu Hackathon-a isprobali smo klasične modele kao SVM koji je dao relativno dobra svojstva, te duboke modele koji su dali značajno bolje rezultate ali su bili teže za trenirati. Svi podaci korišteni su fiktivni i generirani.

## Prijedlog za generiranje pravog skupa podataka:

Iz razloga što VIP nema anotiran skup podataka koji bi podržao ovakav model smislili smo načina kako uz minimalni trošak kreirati kvalitetan skup podataka na temelju neoznačenih podataka. Iskoristili bi podatke o korisnicima koji su u nekom trenutku bili u nekoj od obiteljskih ili grupnih tarifa (savjetovanjem s mentorima smo dobili do znanja da je količina tih podataka i više nego dovoljna za kvalitetno učenje našeg modela).

Ako se tako generiran skup podataka ne pokaže kao dovoljno kvalitetno rješenje sljedeća stvar koja se može napraviti je grupiranje (engl. clustering) korisnika po nekoj mjeri sličnosti gdje se za značajke uzimaju gore navedene značajke. Zatim bi se iz tako generirane skupine ljudi mogao uz preklapanje već označenih korisnika odrediti ostala kućanstva koja nisu bila označena prijašnjim metodama.