TeleSign Datathon

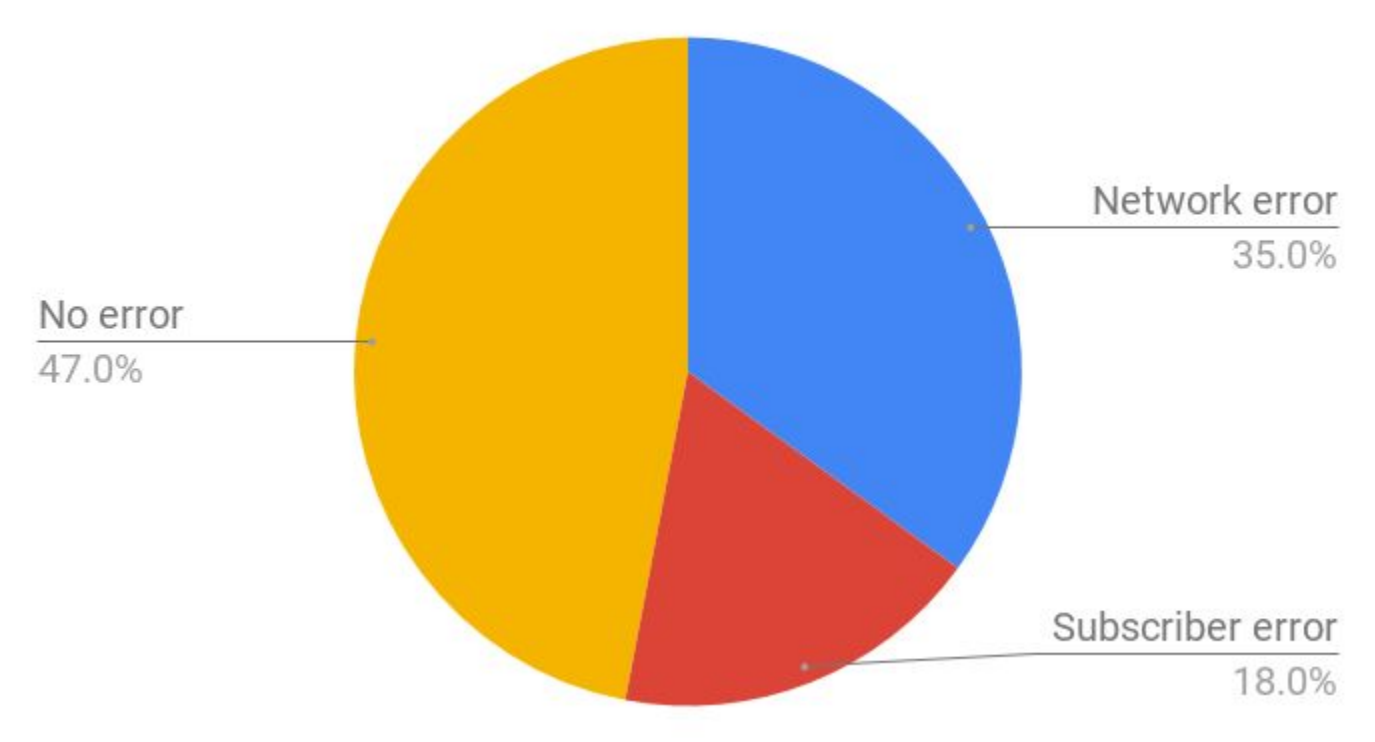
Nikola Jovanović, Marina Ivanović, Petar Veličković, Vladimir Milenković

Feature engineering

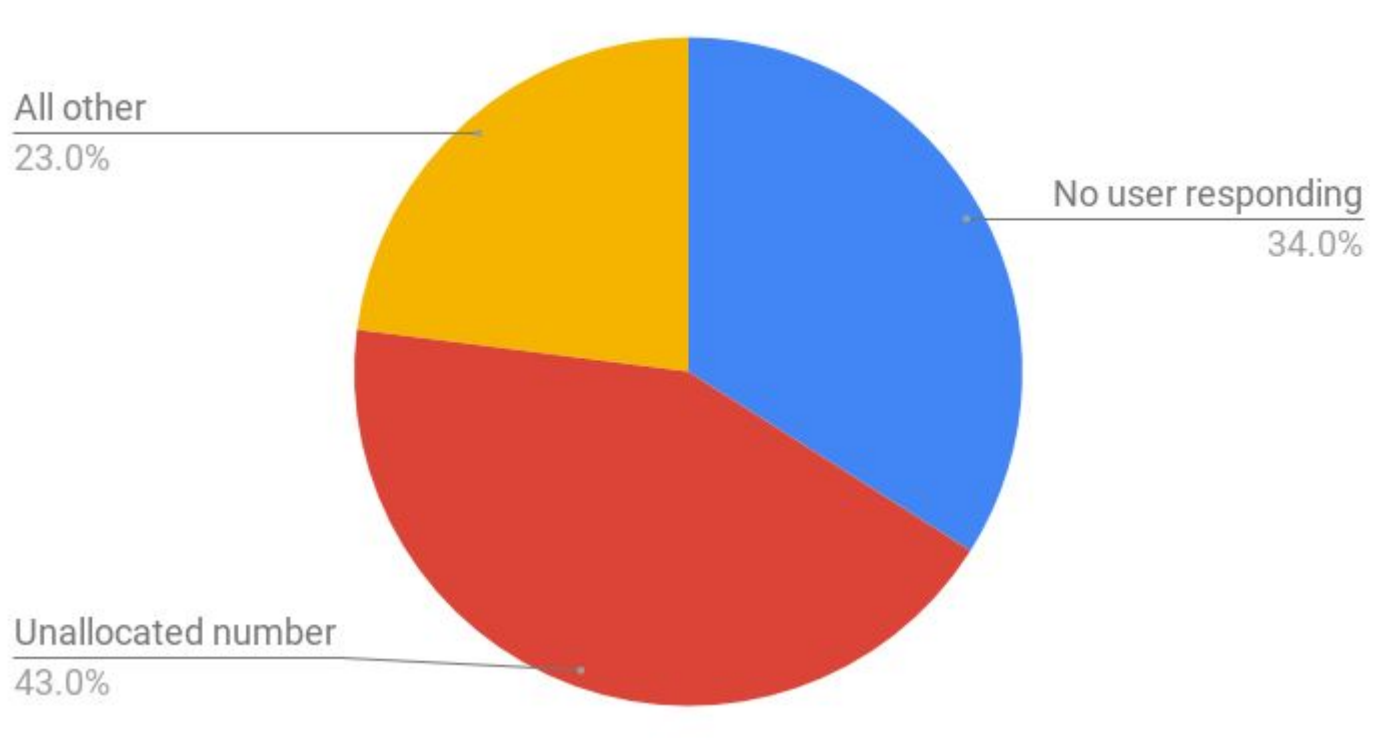
Feature engineering - feature vector

- Za svaki telefonski broj, čuvamo fiksne podatke (državu, blacklist status, itd.).
- Takođe čuvamo i sumarizacije svih poziva u kojima je taj broj učestvovao (odvojeno razmatrajući kada je broj pozivao i bio pozvan).
- Pozive sumarizujemo jednostavnim ručnim pravilima (npr. računanjem srednjih vrednosti).
- => Ušteda na memoriji i jednostavnost pristupa.

Feature engineering - call status



Feature engineering - status name for subscriber error



Feature engineering - roming i burst-ovi

- Poziv je u romingu ako je ispunjen uslov:

  TOC == F/MNO && orig_op != transm_op

- Broj koji kvantifikuje burst-ove za neki telefonski broj:

  % DATETIME intervala između poziva koji su ispodprosečni

Feature engineering - ostali diskretni

- Predstavljeni kao *one-hot* za svaki poziv
- Sumarizovano svojstvo je *srednja vrednost* preko svih poziva

- Primeri:

    - Države u saobraćaju (org, trs, rec, dest)
    - Status poziva
    - Da li je poziv odgovoren?

Feature engineering - kontinualni podaci

- Sumarizovano svojstvo je *srednja vrednost* i *disperzija* preko svih poziva

- Primeri:

    - Interval između poziva (razlika u DateTime vrednostima)
    - Trajanje poziva
    - Vreme uspostavljanja poziva

Ideja

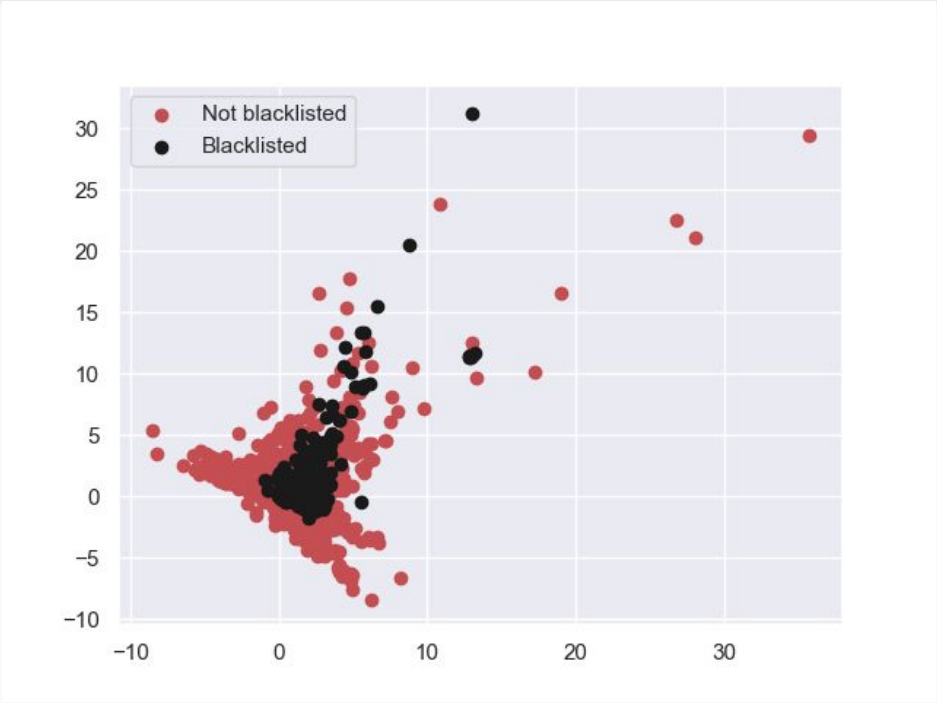
Ideja

- Dataset nije labeliran, stoga pristupamo metodama **nesupervizirano učenja**.
- Tražimo arbitrarnu korisnu *strukturu* u podacima.
- Nakon toga, kombinovanjem interesantne strukture sa nekim karakterističnim svojstvima traženih klasa, moguće je doći do inicijalnog trening skupa za algoritam.

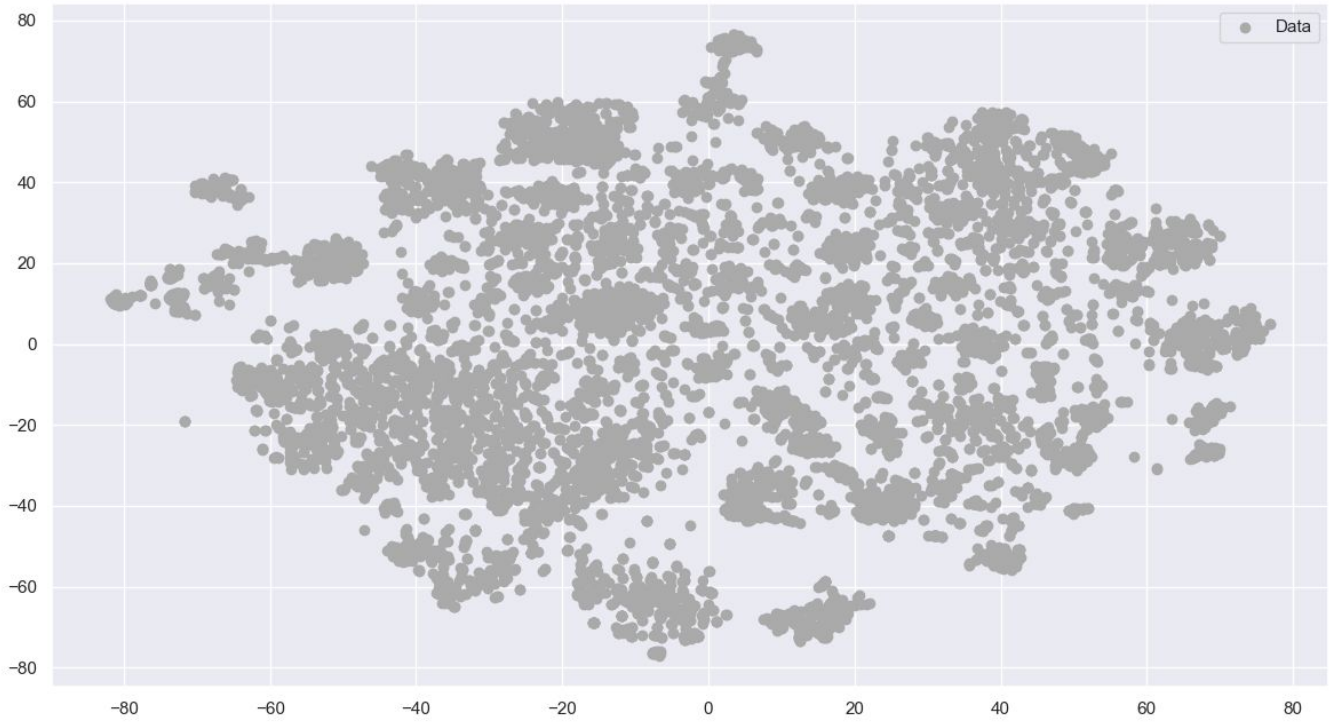
- Labeliranje i verifikaciju labela radimo direktnom analizom podataka u dvodimenzionalnom prostoru.
- => Najpre je potrebno redukovati broj dimenzija u prostoru radi vizualizacije podataka.

Projekcija u prostoru

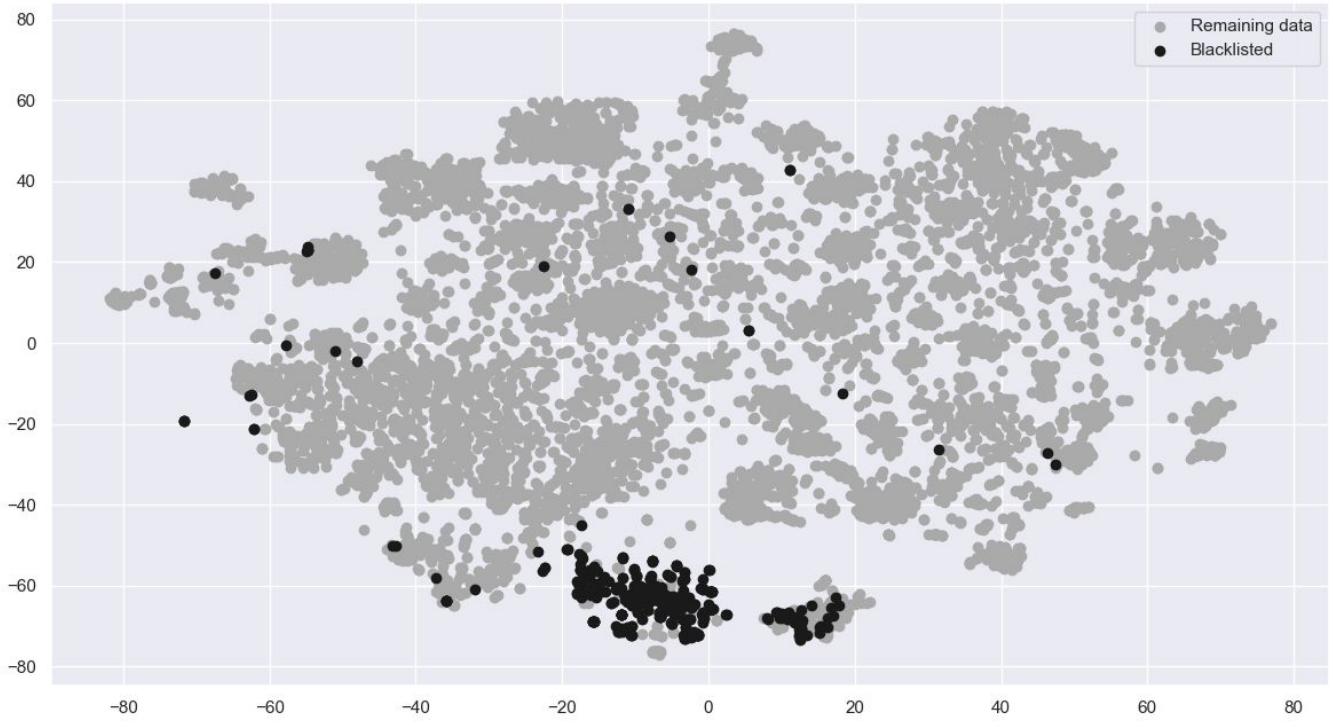
- Isprobane razne opcije za projektovanje u 2D (PCA, Autoenkoderi, t-SNE)
- PCA (desno) i autoenkoderi nisu dali zadovoljavajuće kvalitativne rezultate



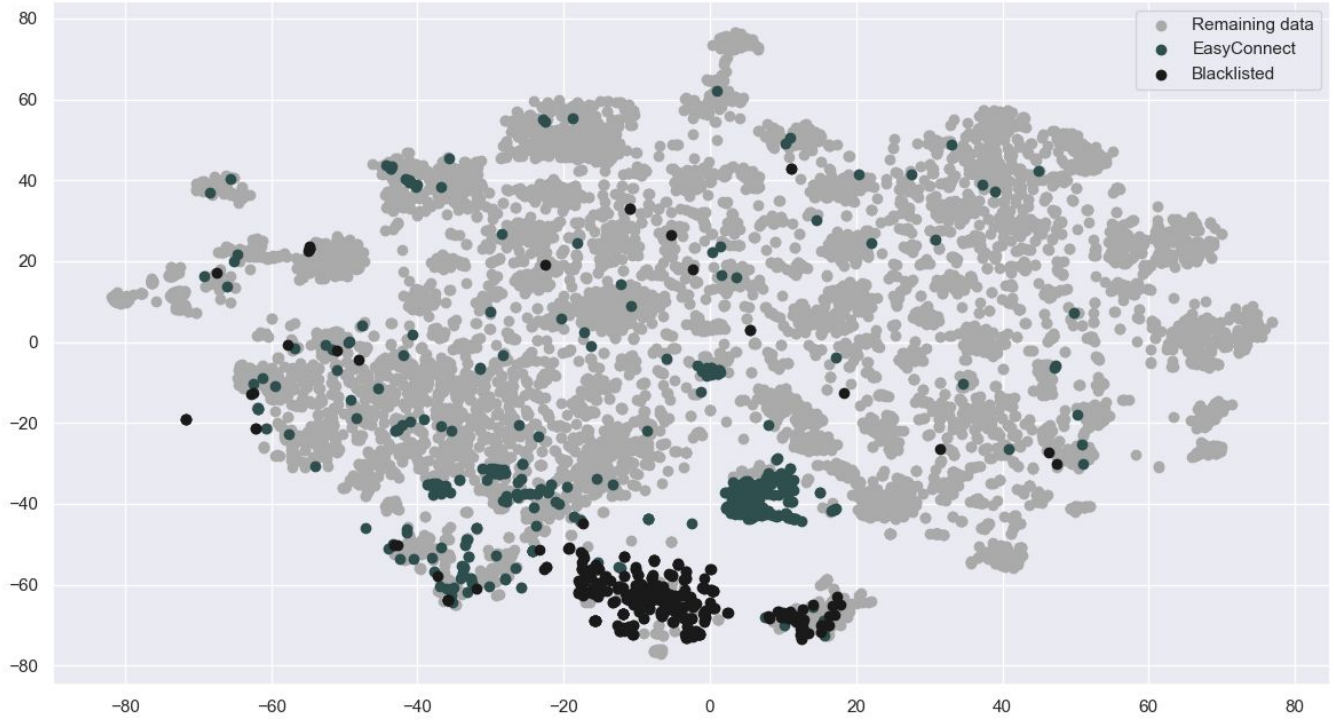
Projekcija u prostoru - t-SNE



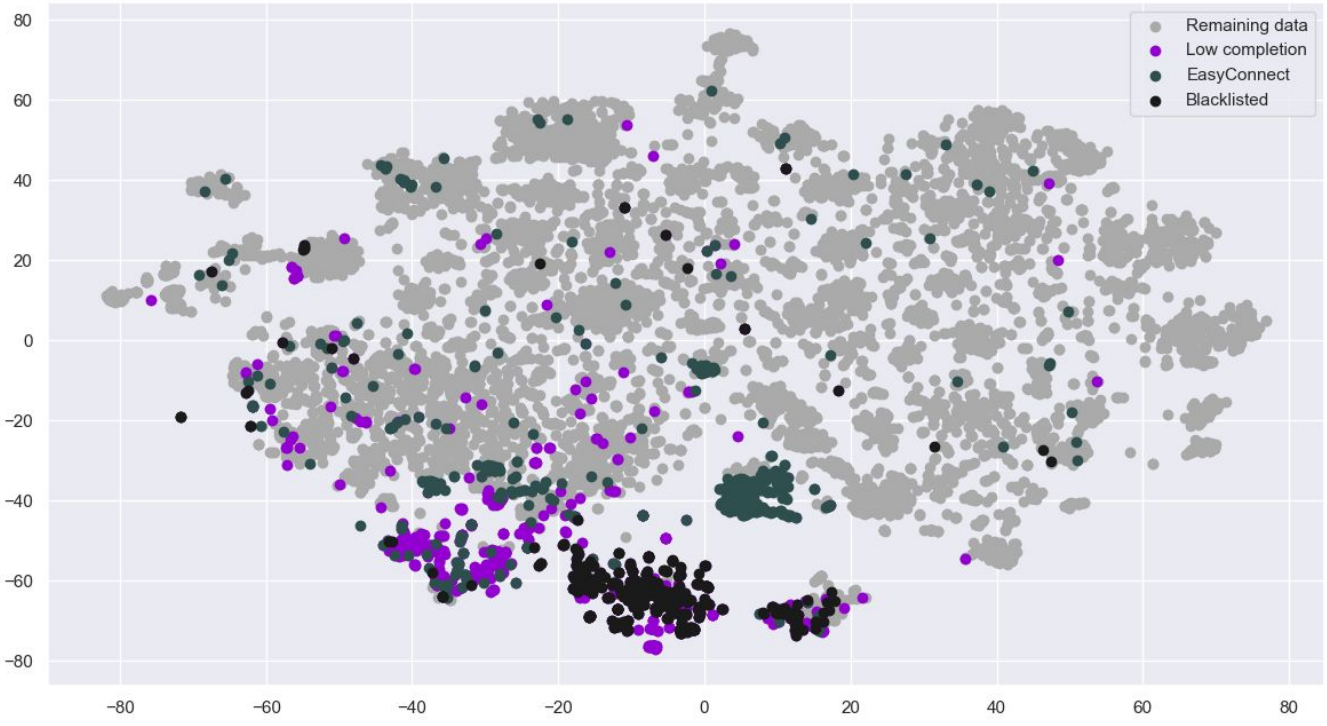
Projekcija u prostoru - t-SNE



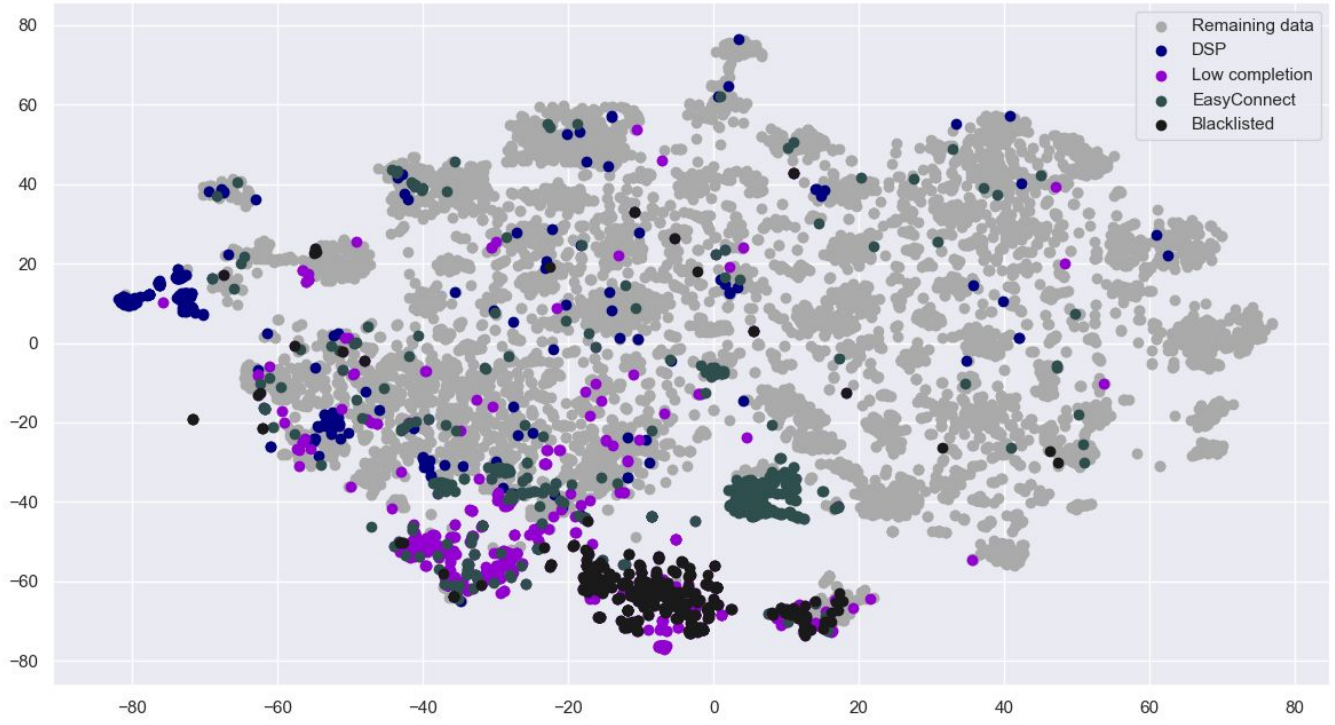
Projekcija u prostoru - t-SNE



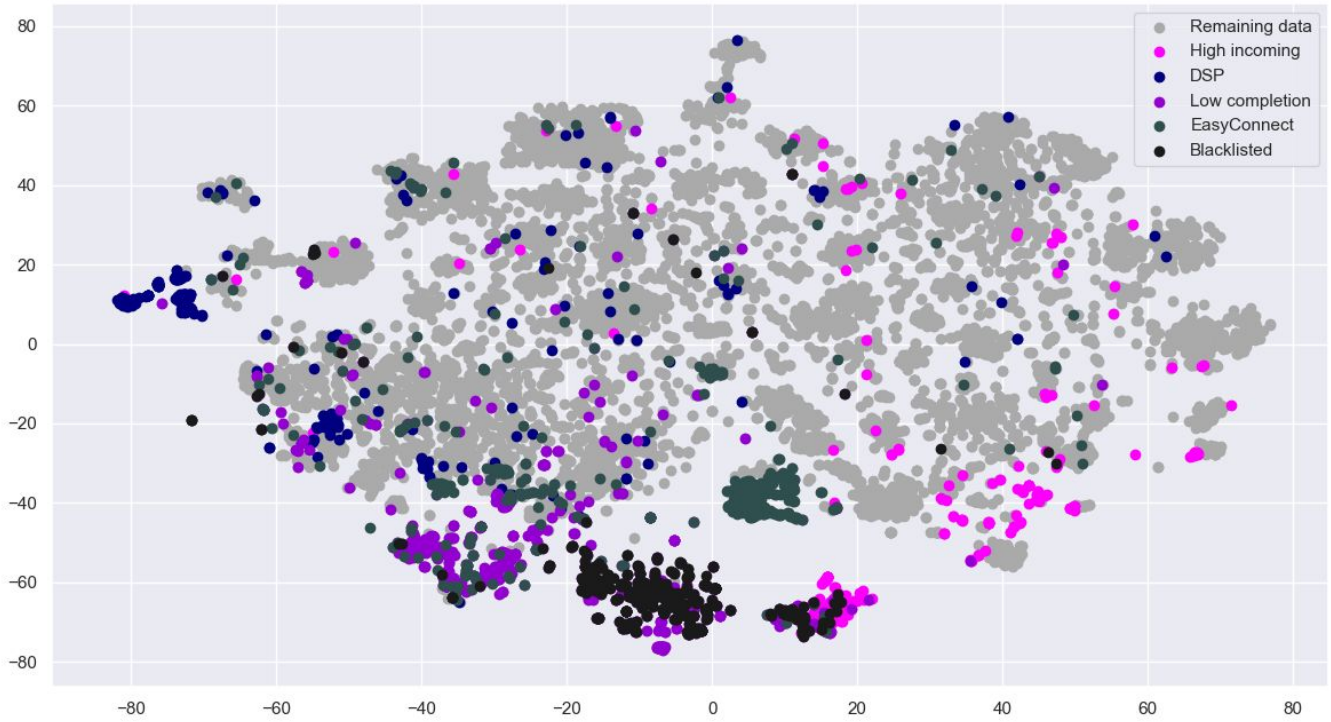
Projekcija u prostoru - t-SNE



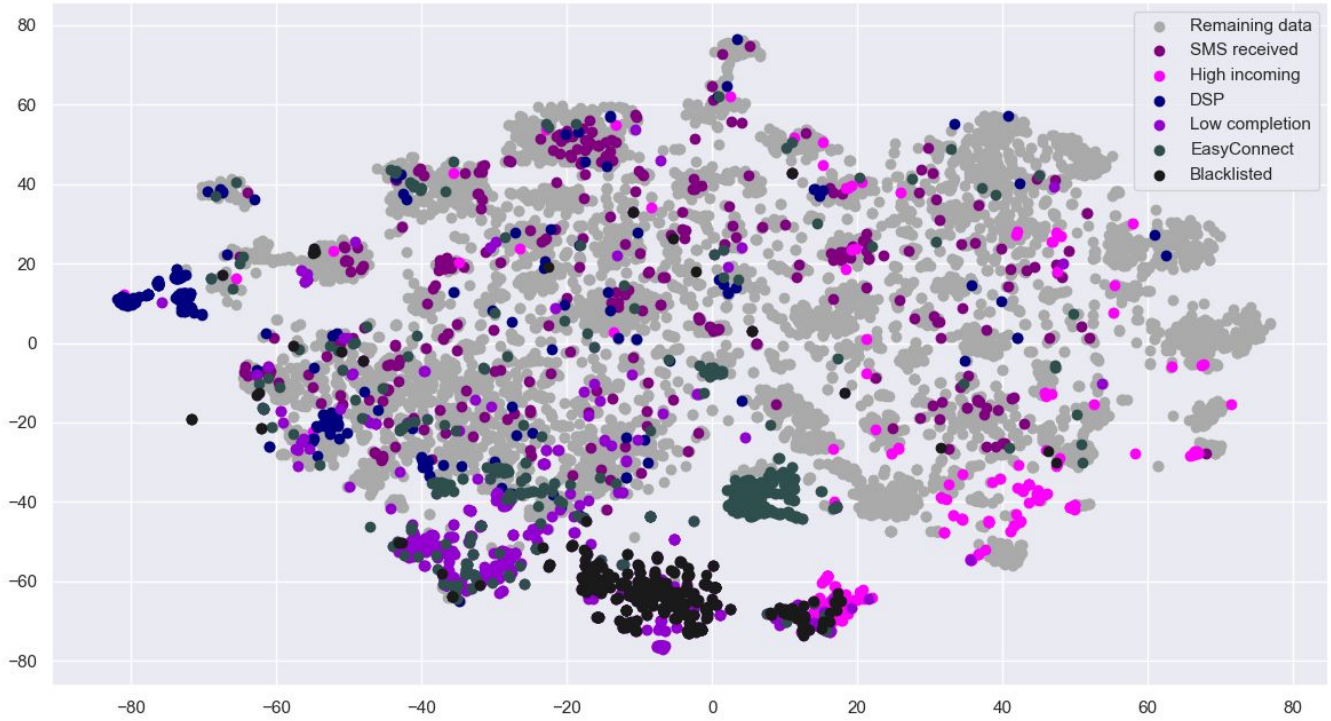
Projekcija u prostoru - t-SNE



Projekcija u prostoru - t-SNE

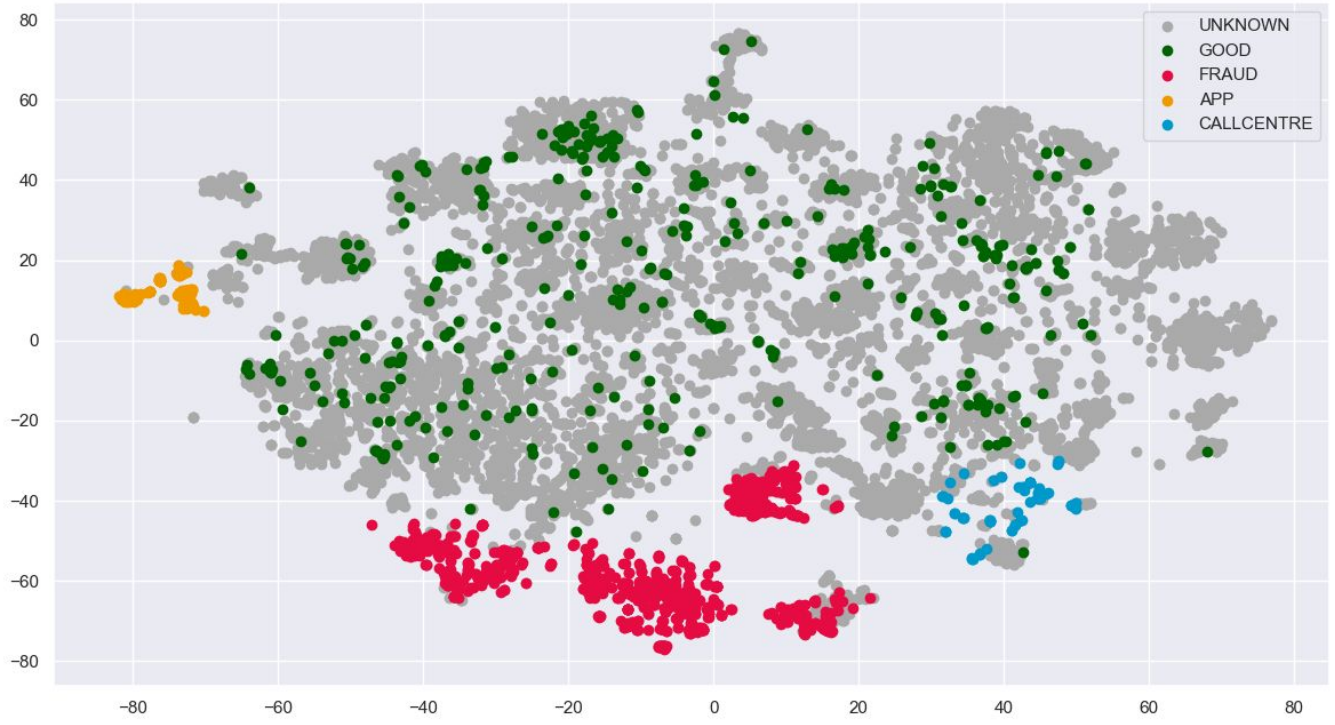


Projekcija u prostoru - t-SNE



Projekcija u prostoru - t-SNE

Izvedene početne labele



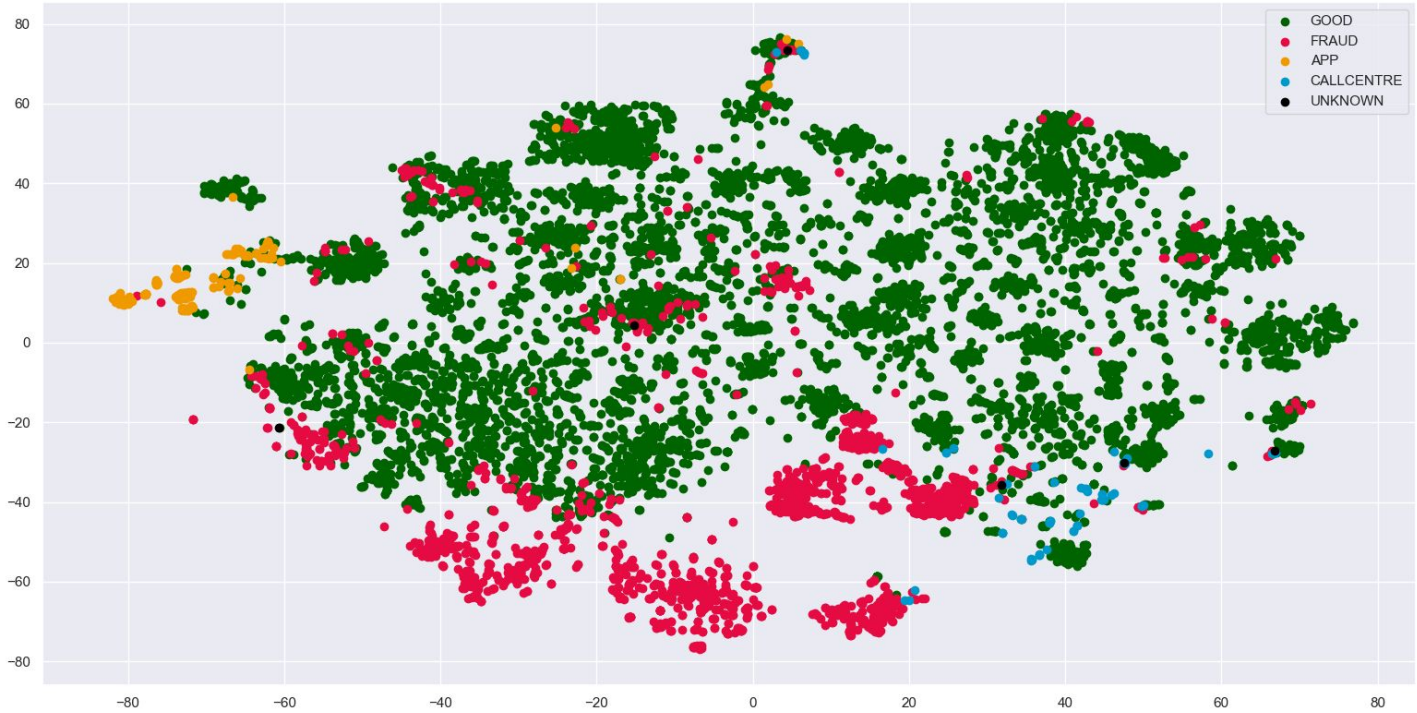
Arhitektura

- Ukupno imamo 2,444 svojstava za svaki broj, koji su izuzetno retko popunjeni (npr. veliki broj je one-hot oznaka za države);
- => Najpre učimo 100 gustih svojstava za svaki broj koristeći autoenkoder.
- Zatim klasifikujemo ova svojstva koristeći jednostavan višeslojni perceptron (MLP) sa tri sloja i ReLU aktivacijama.

Rezultati

- Odvajamo 20% označenih podataka radi validacije
- Višeslojni perceptron postiže **savršenu tačnost** (100%) na ovom skupu.
- Generalizacija na preostale brojeve kvalitativno solidna, sa vrlo malo “nesigurnih” predviđanja (5 primeraka ispod entropijske granice).

Rezultati: predviđanja modela



Diskusija

Algoritam klasifikuje kao *fraudulent* i ulaze koji su daleko od početno identifikovanih skupova!

Identifikovano je da gotovo svi (90+%) ovakvi brojevi iskazuju “sumnjivo” ponašanje, sa bar 10 puta više izlaznih nego ulaznih poziva (ili obrnuto).

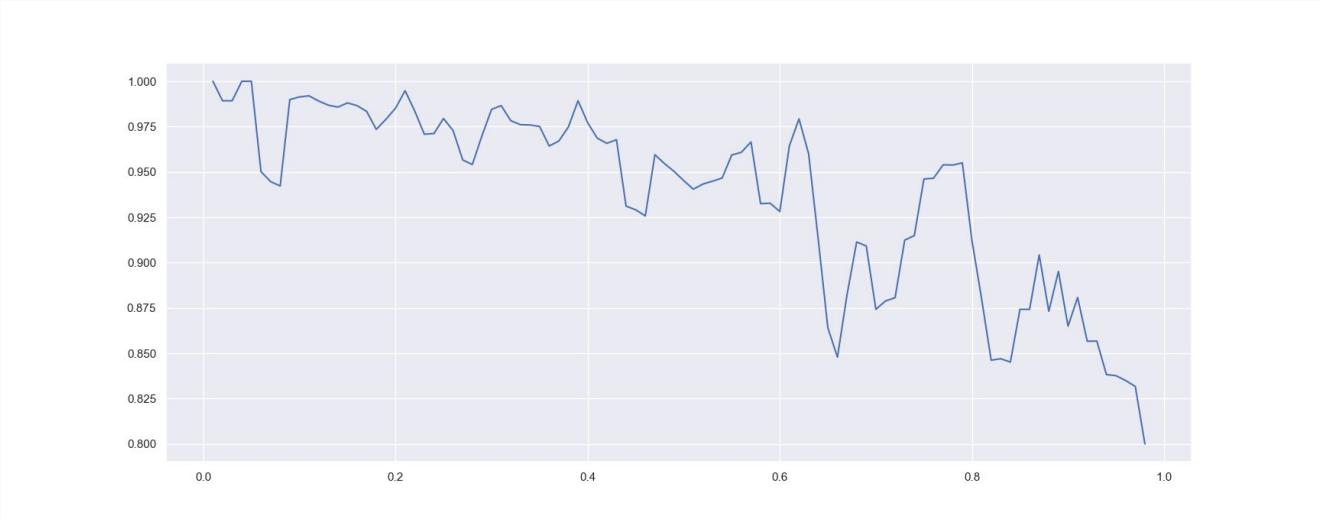
Takođe, preko 45% ovakvih brojeva su iz **Francuske**. (12% u celom datasetu, 27% blacklistovanih)

Skalabilnost metode

- Višeslojni perceptron se primenjuje nezavisno na nivou jednog telefonskog broja, i stoga se trivijalno skalira na datasetove sa arbitrarnim brojem primera.

- Lako je primeniti algoritam na nove brojeve (pošto je *induktivan*), kao i ažurirati sumarizovane podatke (preko pokretnih srednjih vrednosti i disperzija)

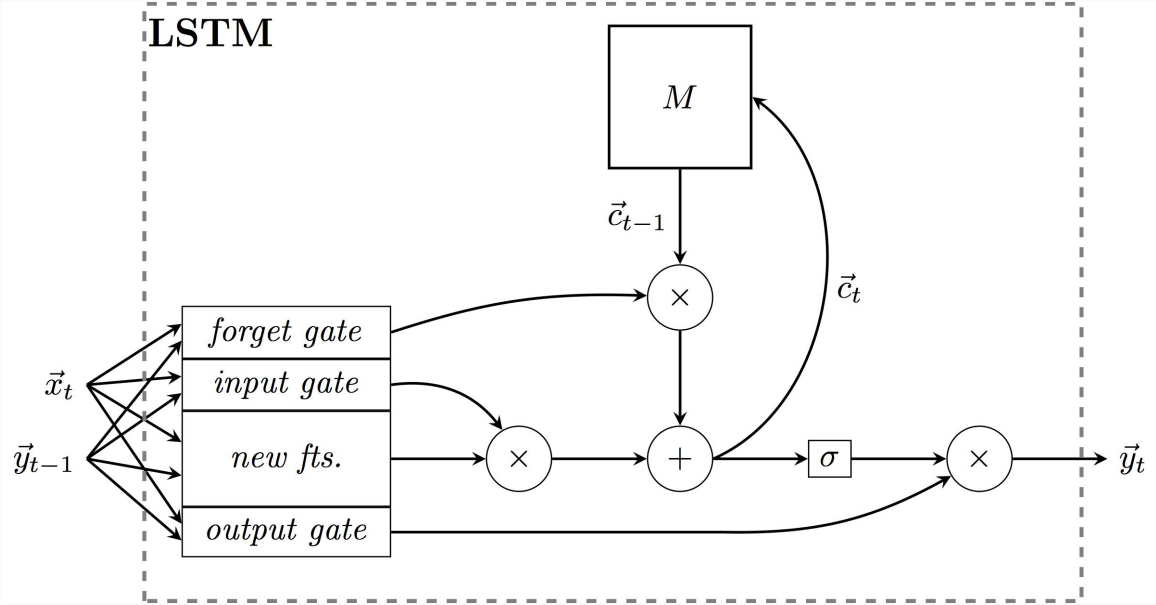
- Neophodan jako mali broj početnih labela za jake rezultate na validacijskom skupu!



Sledeći koraci

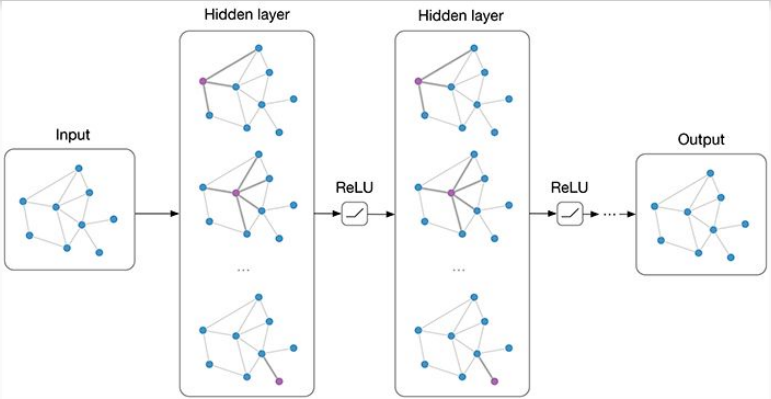
Bolja sumarizacija vremenskih svojstava

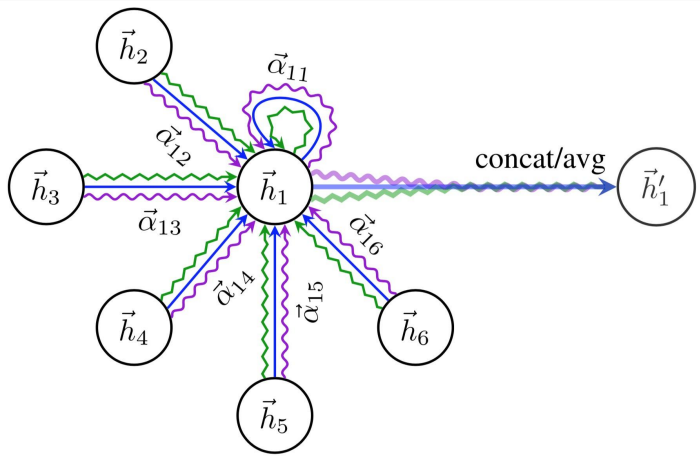
- Koristili smo naivnu sumarizaciju (proseci, prvi momenti itd.) za svojstva individualnih poziva -- ponajviše radi stabilnijeg trening signala i lakših memorijskih zahteva.
- Za precizniju sumarizaciju, moguće je koristiti rekurentne neuralne mreže (npr. LSTM).



Grafovske neuralne mreže

- Moguće je konstruisati graf između brojeva telefona!
- Primenili smo GCN (Kipf & Welling, ICLR 2017), postigli jednake rezultate kao MLP.
- Takođe primenljiv GAT (Veličković et al., ICLR 2018) zbog induktivnih svojstava.
- Moguće skaliranje na velike grafove pomoću metoda poput GraphSAGE (Hamilton et al., NIPS 2017)





Hvala na pažnji!