Andrej Hafner (63160122) **Naloga 3** 6. januar 2019

**Uvod** Izziv je bil izgradnja modela z uporabo linearne regresije, ki bi čim bolj točno napovedala prihode LPP avtobusov v decembru 2012.

**Ocenjevanje točnosti** Za preverjanje točnosti na učnih podatkih sem implementiral prečno preverjanje. Izvajalo se je desetkratno prečno preverjanje, kjer se je 10% podatkov uporabilo za testiranje, 90% pa za učenje. V vsaki iteraciji so se te deli zamenjali zaporedoma, nato pa so se ponovno naučili modeli za posamezne linije. Napovedan je bil čas trajanja vožnje, ki je bil prištet odhodnem času, potem pa je bila izračunana napaka napovedanega časa na podlagi dejanskega časa prihoda. Uporabljena je bila mera za napako mean absolute error.

**Napovedni modeli**

**One hot encoding** Za definiranje atributov iz podatkov, ki predstavljajo določene kategorije, sem uporabil one hot encoding. Tako vsaka možna vrednost dobi svoj stolpec, in če ta atribut zavzame to vrednost ima ta atribut vrednost 1, sicer pa 0. To sem uporabil za predstavitev odhodne postaje, prihodne postaje, smeri vožnje, voznika, ure vožnje in vremenske razmere.

**Krožne vrednosti** Za učinkovito predstavitev minut,ur, dnevov in mesecev sem uporabil sinus in kosinus. Če bi na primer za atribut ura odhoda uporabil vrednosti od 0 do 23, bi to pomenilo da sta 0 in 23 najbolj oddaljena med seboj, čeprav sta zelo zelo skupaj. Ta problem lahko rešimo z dodajanjem dveh atributov, kosinusa normalizirane ure in sinusa normalizirane ure. S tem rešimo problem standardizacije spremenljivke, poleg tega pa je reprezentacija bolj pravilna. Na koncu je bilo to uporabljeno za dneve in mesece, saj se je izkazalo da je za ure bolj učinkovita metoda One hot encoding.

**Vreme** Dodal sem podatke o vremenu iz leta 2012. Uporabil sem preprost One hot encoding za atributa dež in sneg. Izkazalo se je da ni bilo dosti boljše izboljšave rezultate, mislim da zaradi pomanjkanja dnevov, v katerih bi snežilo v učnih podatkih.

Tabela 1: Tabela rezultatov

| metoda | oddaja | ocena - učni | ocena - strežnik | komentar |
|---|---|---|---|---|
| One hot encoding* | prediction13.txt | 150.43 | 180.51 | Najbolj uspešna metoda, dodan še voznik in obe postaji. |
| Krožne vrednosti | predition4.txt | 166.32 | 193.45 | Izboljšanje zaradi dodajanja sinusa in kosinusa. |
| Vreme | predition8.txt | 161.32 | 191.59 | Poslabšanje rezultata za 1 piko po dodajanju vremena. |

**Rezultati**

**Izjava o izdelavi domače naloge.** Domačo nalogo in pripadajoče programe sem izdelal sam.