Počasíčko

Denisa Mensatorisova a Adam Štuller

Obsah projektu je nasledovný. Odovzdávame cely repozitár z github.com. V adresári data sa nachádza spracovaný použitý dataset. Jediny stĺpec, ktorý v ňom nie je spracovaný je LP - v tomto datasete sa nachádza ako dva stĺpce, ktoré sa úplne spracujú až pred analýzov s súbore LP.Rmd. V súbore preprocessing sa nachádza spracovanie datasetu aj ked ten sa už načítava priamo zo súboru. V súboroch, ktoré sa nazývajú ako jednotlivé stĺpce sa nachádza EDA ku každému stĺpcu datasetu. Obe hypotézy overujeme v súboroch hypoteza1 a hypoteza2. Súčasťou odovzdania je aj jedno veľké pdf, kde je všetko spojene.

Je možné, že sa celý repozitár nevmestí do miesta odovzdania kvôli limitu 40 mbi. V takom prípade odovzdáme iba vybranú časť a celý opísaný repozitár sa nachádza v repozitári https://github.com/adamstuller/weather-data-analysis na vetve main.

Dáta obsahujú 413336 pozorovaní nameraných od januára 1973 do decembra 2020. Dataset, ktorý sme si vybrali obsahuje merania zo Sliača, získaných zo stránky organizácie NOAA. Získali sme ich ako objednavku, ktorá je na tejto stránke zadarmo. Detail objednávky (v prípade potreby viete objednavku resubmitnúť na vlastný email) je tu: https://www.ncdc.noaa.gov/cdo-web/orders?email=adam.syn007@gmail.com&id=2545808.

Vzhľadom na to, že obsahuje veľké množstvo atribútov, z ktorých väčšina je pre náš projekt nevyužiteľných, vybrali sme len niekoľko základných atribútov, ktoré sú povinné pre všetky dáta.

Pri skúmaní dát sme zistili, že náš dataset je v podstate časový rad (time series). Práca s časovými radmi je náročná, či už z dôvodu veľkého rozsahu dát (časového obdobia) alebo pre sezónne vplyvy, ktoré sú popísané v ďalších častiach.

Identifikácia problémov v dátach:

V dátach sa nachádza niekoľko chýbajúcich hodnôt, tieto môžu byť nahradené priemerom, alebo lepšou možnosťou mediánom. Avšak v prípade časových radov je ideálne ich nahradiť pomocou interpolácie, ktorá tieto hodnoty nahradí ich odhadom určeným na základe podobných hodnôt v danom intervale.

Okrem chýbajúcich hodnôt dáta tiež obsahujú vychýlené hodnoty, tieto sú ale pre naše hypotézy dôležité nakoľko by ich odstránenie mohlo skresliť výsledky.

V projekte sme použili ako model na predikciu vývoja teploty Naivného Bayesa. Neprogramovali sme ale Bayesovskú sieť. Je to spôsobené podstatou našich dát. Dáta sú jednak časové rady a po párovej analíze sme zistili, že je medzi nimi iba slabá ak vôbec nejaká korelácia. Nebolo teda vhodne tvoriť Bayesovu sieť. Zamerali sme sa preto na iné modely (lineárnu regresiu) ale použili sme aj spomenutého Naivného Bayesa.