Programovanie pre dátovú vedu: projekt 1

September 21- November 26, 2021.
FIIT STU, prednášky: utorok 16:00, cvičenia utorok 18:00, streda 10:00, 12:00 a 17:00.

Očakávaná úroveň vypracovania projektu je len na úrovni prednášok a cvičení, bez hlbšej znalosti vyhodnocovania modelov štatistického učenia.

Termín odovzdania: 1. 11. 2021, 12:00h.
Verzia zadania: 1.02

Očakávaný rozsah práce: Hodnotená bude projektová dokumentácia v rozsahu aspoň troch strán textu s jedným alebo dvoma obrázkami. Samotný text by nemal mať menšiu veľkosť ako 11 bodov. Textovú časť musí sprevádzať spustiteľný kód vo forme R markdown, R notebook alebo R script, pričom zdrojový dátový súbor musí byť pripojený k odovzdaniu, alebo ak to nie je možné vzhľadom na jeho veľkosť, uveďte jeho zdroj tak, aby sme ho moli získať za účelom hodnotenia. Zadania, u ktorých budú chýbať buď zdrojové dáta alebo R kód nebude spustiteľný, budú hodnotené Fx.

Ciele projektu: Cieľom projektu je preukázať schopnosť analyzovať dáta pomocou regresných modelov a testovania nulových hypotéz, za použitia funkcionálneho programovania. Projekt by mal byť prezentovaný ako esej do populárno-vedeckého časopisu v rozsahu maximálne 5 strán, s tým, že si položíte ľubovoľnú otázku a následne sa budete musieť rozhodnúť, čo zo zdrojových dát je použiteľných na to, aby ste otázku zodpovedali. Napríklad, ak by ste analyzovali dáta zmeškaných letov z New Yorku z roku 2013, môžete sa pýtať či existuje nejaký vzťah medzi dĺžkou letu a počtom zmeškaných hodín, alebo či existujú aerolínie alebo letiská ktoré vykazujú väčšiu mieru neskorého odbavovania letov a aké ďalšie parametre sa viažu na tento vzťah. Zamyslite sa, či ide o koreláciu alebo kauzalitu a prečo tomu tak je.

Projekt nemá predpísanú formu vyhotovenia, ale očakávame, že by jadrom bude esej popisujúca nasledovné analytické kroky:

- Nájdite dátový súbor ktorý vás zaujíma. Vzhľadom na charakter očakávaných analýz by bolo vhodné, aby vstupné dáta boli tabelárne s minimálne s desiatkou premenných (angl. variables) a s desiatkami príslušných meraní (angl. values, alebo observations). Vhodným príkladom zdroja takýchto dát je napríklad portál Open Data Bratislava, ODB, (https://opendata.bratislava.sk/), COVID: Our World in Data (https://github.com/owid/covid-19-data/tree/master/public/data), World Bank Open Data (https://data.worldbank.org/), Google Public Data Explorer (https://www.google.com/publicdata/directory), a podobne.
- Stručne predstavte dáta pomocou SQL-LIKE syntaxe a grafiky „grammar of graphics". Môžete uviesť počet záznamov, počet atribútov, ich typy a charakteristiky ktoré sú potrebné k tomu, aby ste dokázali vašu hypotézu. Sú pôvodné atribúty použiteľné, alebo je nutné niektoré transformovať na

nové ? (napríklad ak mám vzdialenosť a čas, môžem si vypočítať rýchlosť a tú použiť pre ďalšiu fázu), majú dáta hodnoty, ktoré významne vybočujú z priemeru? Má dátaset chýbajúce hodnoty? Ako s nimi budete zaobchádzať? Prečo?

- V treťom kroku si stanovte jednu hypotézu, ktoré je overiteľná lineárnym regresným modelom a potvrďte ju pomocou krížovej validácie vstupných dát. Pri programovaní nepoužívajte cykly a špeciálne sa zamerajte na centrálne tendencie a rozptyl parametrov β_1 a β_2.

Forma a hodnotenie odovzdania: Cieľom projektu je osvojiť si základné koncepty funkcionálneho programovania v doméne analýzy dát a štatistického učenia. Vzhľadom na to, že predmet je zaradený do bakalárskeho štúdia, neočakávame od vás vedomosti ohľadne exploratívnej dátovej analýzy ani ďalšie techniky štatistického učenia.

Napriek tomu chceme aby ste preukázali základné zručnosti vo funkcionálnom programovaní a boli schopní vytvárať jednoduché regresné modely na dátach z verejne dostupných zdrojov. Samotné programovanie a analýza dát by vám nemala zabrať viac ako jedno poobedie s tým, že pre vytvorenie projektovej dokumentácie budete potrebovať nejaký čas navyše. Projekt riešite vo dvojiciach podľa vlastného výberu. Trojice nie sú povolené. Projekt musí obsahovať tiež referenciu na dátovú sadu použitú na vytvorenie modelu – ideálne ako *tsv/csv* súbor pripojený k odovzdaniu, respektívne ako link na verejne dostupné zdroje. Je dovolené aby viaceré dvojice pracovali na rovnakom datasete, ak ich hypotézy a regresný model budú odlišné.

V texte musíte preukázať schopnosť popísať podstatné časti vašej analýzy tak, aby ste presvedčili laika, ktorý nevidel váš kód ani nevie programovať, ale zároveň musí byť dodatočné detailná, aby presvedčila experta, ktorého zaujíma váš kód a dáta. Podanie nekomentovaného zdrojového kódu s grafmi, ktoré nemajú popísané osy, nie je jasné prečo boli vybraté a čo zobrazujú, bude hodnotené známkou Fx.

Hodntiace kritérá: Váš projekt bude hodnotený s ohľadom na štyri kritériá – prvé a druhé nesú mierne väčšiu váhu ako tretie a štvrté.

- Ste schopný efektívne funkcionálne programovať pre potreby dátovej analýzy v programovacom prostredí R?
- Preukázali ste schopnosť aplikovať lineárne regresné modely a testovanie nulových hypotéz tak, aby ste poskytli presvedčivý dôkaz pre zvolenú hypotézu? Chápete koncept modelu, viete ho správne interpretovať a viete dokázať že je stabilný?
- Ste schopný prezentovať vaše argumenty v prospech zvolenej hypotézy stručne a jasne?
- Je grafické vyhotovenie a priložený kód dostatočne dokumentovaný? Sú grafy vhodne vybraté a anotované?