Data Science (aka Regionális gazdaságtan)

A kurzus célja, hogy a diákokat - statisztikai előképzettségükre építve - bevezesse a modern adatelemzés (data science) rejtelmeibe. Valós adatokra és konkrét üzleti alkalmazásokra építve, aktív órai munkával igyekszünk bemutatni a data scientist napi munkájához szükséges eszköztárat. A tanultakat a diákok egy maguk által választott témában, önálló projekt keretében hasznosítják.

A kurzushoz kapcsolódó anyagok a kurzus github oldalára kerülnek ki.

Előfeltétel

A kurzus feltételezi a `git`, illetve az `R` és az `RStudio` alapvető ismeretét. Ennek megszerzéséhez a nyár elejére összeállítottunk egy segédanyagot, melyet a kurzus github oldalára feltöltött a "prerequisite" mappába gyűjtöttünk. Az alapismeretek feltétlen szükségesek a kurzus elvégzéséhez. Az alapok meglétét az első órán esedékes beugró teszt sikeres teljesítésével kell bizonyítani.

Oktatás kerete

A kurzus időkerete hetente 2x90 perc. Minden alkalom egy elméleti és egy gyakorlati részből áll. Az elméleti rész is igyekszik gyakorlati lenni, a mindennapi munkánkból vett példák és valós esetek bemutatásával. A gyakorlati rész alatt a diákok saját gépeiken programozva önállóan alkalmazhatják az előző rész során bemutatott módszereket (óránként két oktató segítségével).

Oktatók

Bérczi András, Divényi János, Holler Zsuzsa, Kocsis Gábor, Koncz Tamás, Lukács Péter - az Emarsys data science csapatának tagjai

Kontakt: eltecon.ds kukac gmail pont com

Értékelés

Az évvégi jegy két részből tevődik össze: kisebb részben a heti házi feladatok (30%), nagyobb részben a félévi projekt (70%) alkotják. A házi feladatok nagy része a projekt előkészítéséhez kötődik, amelyet a vizsgaidőszakban egy később megbeszélt alkalmon kell majd előadni.

Hasznos anyagok

- James - Witten - Hastie - Tibshirani: An Introduction to Statistical Learning
- Grolemund - Hadley: R for Data Science
- Gentzkow - Shapiro: Code and Data for the Social Sciences: A Practitioner's Guide
- (Emarsys Craftlab blog - data section](https://blog.craftlab.hu/tagged/emarsys-data)

Tematika

1. **szeptember 11.** Bevezetés: reprodukálhatóság, verziókövetés Git-tel, R projekt setup, mappastruktúra, hasznos segédfájlok. (*Kocsis Gábor*)
2. **szeptember 18.** Adatgyűjtés, feltáró adatelemzés (Exploratory Data Analysis), adattisztítás. (*Divényi János*)
3. **szeptember 25.** Adatvizualizáció: grammar of graphics, ggplot2, irányelvek. (*Koncz Tamás*)
4. **október 2.** Kutatási riport (RMarkdown), interaktív vizualizáció (plotly). (*Koncz Tamás*)
5. **október 9.** Hatásmérés kísérletezéssel: feltételek, minimum detectable effect. (*Bérczi András*)
6. **október 16.** Mérés bizonytalansága, szimulációs módszerek: Monte-Carlo, bootstrap, permutációs teszt. (*Bérczi András*)
7. **október 23.** TANÍTÁSI SZÜNET
8. **október 30.** TANÍTÁSI SZÜNET
9. **november 6.** Gépi tanulás (machine learning) bevezetés. Supervised learning alapok: lineáris és logisztikus regresszió. Predikció versus okság. Előrejelzési pontosság versus interpretálhatóság. (*Lukács Péter*)
10. **november 13.** Előrejelzési pontosság mérése, modellkiválasztás. Torzítottság versus variancia (bias-variance trade-off). Regularizáció. Cross-validation. (*Holler Zsuzsa*)
11. **november 20.** Supervised learning modellek a linearitáson túl: shrinkage regresszió (LASSO), döntési fa. (*Holler Zsuzsa*)
12. **november 27.** Előrejelzési pontosság javítása mintavételezéssel: boosting és bagging modellek, model ensemble. (*Divényi János*)
13. **december 4.** Unsupervised learning: klaszterezés (K-means, hierarchical), dimenzió redukció (PCA). (*Kocsis Gábor*)
14. **december 11.** Projektmunka folyamatos konzultációs lehetőséggel. (*az összes oktató részvételével*)