Kivonat

Automatizált változó generálás

Kubicza Gréta Andrea

Témavezető: Dr. Kovács Edith Alice

A szakdolgozatomban a Python egy új könyvtárának, az `autofeat`-nek a megismerésével foglalkoztam. Ez a könyvtár a meglévő magyarázóváltozókból generál új változókat. Az eljárás a kiinduló változók bizonyos nemlineáris transzformációit is beépíti a változók közé, majd azokat tartja meg, amelyek relevánsak a célváltozókra. Ha ezeket lineáris, illetve logisztikus regresszióval kombinálva használjuk, jól értelmezhető és magyarázható modelleket kapunk. Hátránya, hogy nagyon kibővíti a magyarázóváltozók halmazát, ezzel megnövelve a futási időt is, valamint kisméretű adathalmazok esetén előfordul a túltanulás.

Ezen felül megvizsgáltam, hogy érdemes-e dimenziócsökkentést alkalmazni az új változók generálása előtt. Két eljárást vizsgáltam meg a Random Forests Feature importance-t és a Boruta algoritmust, amelyek a változók fontosságnak segítségével végeznek dimenziócsökkentést. Részletesen tanulmányoztam több regressziós és osztályozós modellt (Random Forests, Support Vector Machines), azoknak a matematikai hátterét, valamint, hogy milyen eredményeket érek el, ha az `autofeat`-tel együtt alkalmazom őket.