Utazási élmény prediktálása

Az adatállomány azon utasok véleményéről tárol adatokat, akik valaha utaztak Shinkansen vonaton. Több különféle faktort figyelembe vesz, amely befolyásolhatja az utazás minőségét, például az ülés kényelmessége, a vonat pontossága, sőt, hogy milyen könnyen található meg az indulási vágány. A végső elégedettségi szintet az Overall\_Experience attribútum mutatja, ahol 0 az érték, ha az utas összességében nem elégedett a szolgáltatással, 1 pedig, ha igen. Ezeket az adatokat a Surveydata\_train állományban találjuk. Emellett az utasokról is tárol információkat a Traveldata fájlokban, mint például az utas nemét, korát, utazásának hosszát. A modellillesztések során ezeket az adatokat is figyelembe vettük.

Arra a kérdésre kerestük a választ, hogy a kapott adatok birtokában melyik modell képes pontosabb eredmény prediktálására. De először is, ehhez szükségünk volt egy kis előfeldolgozásra. A Surveydata\_train és Traveldata\_train fájlokkal dolgoztunk, mert felügyelt tanítással szerettük volna megoldani a feladatot. A két CSV fájl beolvasása után mindkét állományon beállítottuk az ID oszlopot id szerepűnek, az Overall\_Experience oszlopot pedig labelnek. Id alapján összekapcsoltuk a két állományt, így már minden attribútumot vizsgálhattunk egyszerre. Mappel átalakítottuk a vélemények szöveges változatát (good, manageable, poor, stb) számokra az egyszerűbb kezelhetőség érdekében. Végül az Overall\_Experience attribútumot alakítottuk át, mert integer típusúnak ismerte fel, de nekünk binomiális típusra volt szükségünk. Az előfeldolgozás végén letároljuk az adatállomány a Ai Studio bináris fájlformátumában, illetve kiengedjük a kimenetre.

Az adatok előkészítése ezzel befejeződött, kezdhetjük a modellek illesztését. Ezt egy másik fájlban valósítottunk meg. Három modellt illesztettünk az adatokra: neurális hálót, döntési fát és logisztikus regressziót. Mielőtt odaadtuk volna az állományt a modelleknek, előtte két részre osztottuk: train és test adatokra. A neurális háló alkalmazása előtt MISSING értékkel helyettesítettük a hiányzó adatokat és az összes nominális értéket számra alakítottuk át. Erre nincs szükség a további két modellnél, csupán itt. A modell 20 cikluson keresztül dolgozik, 0,01 learning rate-tel és 0,8-as momentummal. A hálónak 2 rejtett rétege van, az első 8, a második 4 node-ot tartalmaz. Tapasztalataink alapján ez jó eredményt adott. A performance-ot binomiális performanciára nézzük minden esetben, amely a neurális háló esetén 94,4 % körül ingadozik (különböző seedektől függően). A ROC görbéje itt látható:

A képen szöveg, sor, Diagram, képernyőkép látható

Automatikusan generált leírás

Ezt követően egy döntési fát illesztettünk az adatokra. Itt a cél az volt, hogy megtudjuk, milyen mélységig érdemes felépíteni a fát, így ezt a paramétert egy változóban módosítjuk. 2-es mélységtől, egészen a 12-es mélységig próbálgatja a program a modellek illesztését. Ahogy az ábrán látható, 8-as mélységig látványosan nő a pontosság, ezt követően csupán enyhe javulás tapasztalható. Úgy gondoljuk, hogy a 8-as mélységű modellt érdemes használni, mivel az 91,0 %-os eredményt ért el. A legjobban a 12-es mélységgel működött, ez 92,0 %-os pontosságú. A kisebb méretű fák a Seat\_comfort és Onboard\_experience attribútumokat használják, a nagyobbak már a többit is bevonják a kategorizálásba. Az említett 2 attribútum vélhetően fontos lehet a predikció szempontjából.

A képen szöveg, szám, sor, képernyőkép látható

Automatikusan generált leírás

Végül egy logisztikus regressziót illesztünk az adatokra. Az IRLSM solvert használjuk, a többi értéket alapértelmezetten hagyjuk, ez 90,23 %-os pontosságot eredményez.

Összességében a kipróbált modellek jól prediktálják az összesített utazási élményt, mindegyik 90 % felett teljesít. A legjobban a neurális háló teljesített, ami eléri a 94 %-ot is. Kérdéses viszont, hogy mennyi erőforrást szeretnénk a prediktálásra szánni, mivel a neurális háló kifejezetten erőforrásigényes tud lenni. Utána a döntési fa a maga 91-92 %-os eredményével, amelynél tudjuk szabályozni a mélységet. Nem feltétlenül érdemes a 12-es mélységet használni, mert nem javítja már jelentősen az elért eredményeket. A legjobban a 8-assal járunk, ez adja a legjobb pontosságot az erőforrásfelhasználás függvényében. Ezután nem sokkal lemaradva a logisztikus regresszió jön pontosságban. Ha a többi modell nem nyerte el a tetszésünket, vagy esetleg túl sokáig tartana a futtatásuk, ezzel járunk a legjobban.