Automatické generovanie prediktorov zo semi-štrukturovaných dát

Základom je opäť platforma 7Segments, ktorá obsahuje rôznorodé dáta o zákazníkoch, ktoré reprezentujú rôzne udalosti („eventy") týkajúce sa jednotlivých zákazníkov. Napr. zákazník urobil takúto operáciu, zákazník zakúpil takýto produkt, zákazník ma aktivované takéto funkcie, atď. 7Segments má k dispozícii rôzne datasety.

Cieľom je vytvoriť metódu, ktoré budú schopné z týchto dát identifikovať/etrahovať kľúčové atribúty („prediktory") ovplyvňujúce zákazníkom zvolený cieľový atribút (numerická, nominálna alebo binárna premenná). Relevantnosť a sila vplyvu medzi „prediktorom" a cieľovým atribútom môže byť vyhodnotená pomocou:

- R^2 / korelačný koeficient pre numerickú premennú .
- ChiSq / pravdepodobnosť zamietnutia H0 pre nominálnu alebo binárnu premennú.
- WoE / Information Value pre binárnu premennú.

Extrahované „prediktory" budú následne použité ako vstupné atribúty pre rôzne algoritmy strojového učenia (rozhodovacie stromy, regresia, neurónová sieť, SVM), ktoré sa používajú ako bežné podporné riešenia pre úlohy marketingového typu. Výsledkom budú napr. klasifikačné alebo predikčné modely.
Príklad: „prediktor" = suma výšky nákupov v kategórii "topánky" za posledne 3 mesiace.

Pri extrakcií ďalších „prediktorov" je nutné vziať do úvahy ich koreláciu s už existujúcimi „prediktormi" (napr. max. výška nákupov koreluje s výškou nákupov) a takisto preferovať také „prediktory", ktoré budú zvyšovať celkovú silu modelu, v rámci ktorého budú použité ako vstupy.

Celkovo je cieľom, aby bolo možné vytvoriť čo najlepší model pozostávajúci z extrahovaných „prediktorov" s minimálnym počtom pokusov.
Príklad: celkovým výsledkom je klasifikačný model reprezentovaný rozhodovacím stromom, ktorý je schopný rozdeliť zákazníkov do dvoch skupín (s čo najvyššou presnosťou), čiže buď verný zákazník alebo zákazník odíde. A tento rozhodovací strom je práve zložený z identifikovaných prediktorov.

Odporúčané technológie: Python, C++, C# alebo Javu.

Riešenie tejto úlohy pozostáva z:

- Analýza a pochopenie dát, ktoré majú byť základom pre extrakciu „prediktorov".
- Návrh algoritmu pre automatickú extrakciu „prediktorov".
- Testovanie kvality extrahovaných „prediktorov" prostredníctvom tradičných metód strojového učenia.