

# AUTOMATICKÉ GENEROVANIE PREDIKTOROV ZO SEMI- ŠTRUKTÚROVANÝCH DÁT

Prezentácia k riešeniu čiastkovej úlohy  
UVP Technicom – Aktivita 3.1 – PP4

Michal Puheim

# Cieľ

- ▣ Automatická identifikácia/ generovanie takých kľúčových atribútov (*prediktorov*) z databázy údajov o zákazníkoch, ktoré majú vplyv na určený cieľový atribút.
- ▣ Získané *prediktory* budú použité pre rozhodovacie a predikčné algoritmy strojového učenia v marketingových úlohách.
- ▣ Príklad: *prediktor* = suma nákupu za 3 mesiace.

# Postup riešenia úlohy

- ▣ Analýza a pochopenie východzích údajov.
- ▣ Návrh a analýza algoritmu pre automatickú extrakciu *prediktorov*.
- ▣ Testovanie kvality *prediktorov* v algoritmoch strojového učenia.

# Údaje

- ▣ Údaje o zákazníkoch môžu byť dvoch typov:
  - atribúty – vždy jedna hodnota
  - udalosti – určené viacerými hodnotami (atribútmi)
  
- ▣ Hodnoty atribútov môžu byť:
  - binárne
  - nominálne
  - numerické
    - ▣ *celé čísla*
    - ▣ *reálne čísla*

# Atribúty

- ▣ Hodnoty priradené konkrétnemu zákazníkovi.
- ▣ Zákazník je určený identifikátorom *customer\_id*.

customer_id	properties.first_name	properties.last_name	properties.gender	properties.birth_year
53039a0c25030f78d9d4dbfd	Katarina	Petrikova	female	1960
53039a0c25030f78d9d4dc01	Božena	Rendlová	female	
53039a0c25030f78d9d4dc05	Mariana	Šutá	female	1972
53039a0c25030f78d9d4dc09	Stanislav	Budzinak	male	
53039a0c25030f78d9d4dc0a	Rudolf	Adamov	male	1972
53039a0c25030f78d9d4dc0e	František	Halčín	male	1959
53039a0c25030f78d9d4dc10	Meno	Nezadané		
53039a0c25030f78d9d4dc14	Janka	Beniačová	female	1966
53039a0c25030f78d9d4dc18	Jana	Briešková	female	1982
53039a0c25030f78d9d4dc1d	Marie	Blaňárová		
53039a0c25030f78d9d4dc23	Zuzana	Kocianová	female	
53039a0c25030f78d9d4dc27	Erika	Bodová	female	1986

# Udalosti

- ▣ Skupiny atribútov priradené zákazníkovi.
- ▣ Ich počet je variabilný.
- ▣ Každý typ udalosti má vlastné atribúty.


customer_id	events.purchase. category_id	events.purchase. category	events.purchase. product_id	events.purchase. product	events.purchase. count	events.purchase. profit
53039a0c25030f78d9d4dbfd	1017	Shoes	122941	AUTHORITY-Kala	1	5.98
53039a0c25030f78d9d4dc01	1017	Shoes	112049	ADIDAS-SE Daily Vulc	1	10.65
53039a0c25030f78d9d4dc05	1001	accessoryapparel AAC	108737	EXISPORT-F GLOVES UNI	1	1.58
53039a0c25030f78d9d4dc05	1012	football ITG	113814	NIKE-TIEMPO NATURAL IV IC	1	6.84
53039a0c25030f78d9d4dc05	1014	skiingsnowboardin g OSS	115245	AUTHORITY- GRETTE II	1	3.75
53039a0c25030f78d9d4dc05	1022	homefitness IFI	108174	KETTLER-GYM BALL 65 cm green	1	8.97
53039a0c25030f78d9d4dc05	1032	tennis ITG	134608	WILSON-BLX Tour Backpack BKGD	1	9.14

# Relevantnosť prediktora

- ▣ Určenie závislosti (sily vplyvu) medzi prediktorom a cieľovým atribútom.
  
- ▣ Môže byť vyhodnotená:
  - Korelačným koeficientom
    - ▣ pre numerické atribúty,
  - Chi-kvadrát testom nezávislosti
    - ▣ pre binárne a nominálne atribúty.

# Typ atribútu

- ▣ Pre výber spôsobu vyhodnotenia relevantnosti *prediktora* je nevyhnutné najprv určiť údajový typ atribútu:

- binárny,
  - nominálny,
  - numerický,
    - ▣ celé číslo,
    - ▣ reálne číslo,
- 
- nominálny (diskrétny ),
    - ▣ chi-kvadrát test nezávislosti,
  - numerický (spojitý),
    - ▣ korelačný koeficient.



# Generátor prediktorov

- ▣ Pre numerické atribúty:
  - aritmetické operácie,
  - agregáčné operácie ,
  - inšpirácia z teórie riadenia.
  
- ▣ Pre nominálne atribúty:
  - logické operácie, pravidlové odvozovanie,
  - viachodnotová logika,
  - početnosť (pri udalostiach).

# Overenie a testovanie

- ▣ A) Implementácia vlastných algoritmov strojového učenia pre účely testovania.
- ▣ B) Nahratie vytvorených prediktorov do databázy klienta a testovanie pomocou už navrhnutých prostriedkov.

# Zhrnutie

## ▣ DONE

- Analýza, pochopenie údajov.
- Získanie údajov z databázy klienta.

## ▣ WIP

- Načítanie údajov do pamäte programu.
- Určenie dátových typov údajov.

## ▣ TODO

- Implementácia overenia relevantnosti *prediktora* voči cieľovému atribútu.
- Vytvorenie generátora nových *prediktorov*.
- Overenie a testovanie.

Ďakujem za pozornosť!