Projekt

M8DM1 - Data mining I

Simon Sudora

Zadanie

Nová televizní stanice chce vstoupit na český trh. Jak by se měla profilovat, aby odpovídala častým a typickým potřebám české populace?

Úvod

Ak chce nová televízna stanica na českom trhu uspieť, je potrebné nájsť **dieru na trhu**. To inak povedané znamená, že musí naplniť svoje vysielanie obsahom, o ktorý majú ľudia záujem, ale zatiaľ nie je dostatočne pokrytý existujúcimi stanicami.
Stanice môžeme na základe nimi vysielaného obsahu rozdeliť do nasledovných kategórií:

- **Všeobecné**
  Obsah ktorým napĺňajú vysielanie je veľmi rôzny - od rozprávok pre deti cez spravodajstvo až po reality show. Vysielaný **obsah je veľmi flexibilný** - môžu jednoducho reagovať na aktuálne požiadavky divákov. Tento typ staníc sa zameriava na **masy**, a preto ich hlavným zdrojom príjmov sú **reklamy**.
- **Úzko profilované**
  Tieto stanice sú spravidla zamerané na iba j**eden typ obsahu**. Na rozdiel od všeobecných staníc sa zameriavajú iba na **úzku skupinu ľudí** s určitým záujmom, ktorí sú za kvalitný obsah ochotný viac platiť.
  Možné oblasti zamerania:
  - Športové
  - Dokumentárne
  - Filmové/Seriálové
  - Erotické
  - Hudobno-lifestylové
  - Detské
  - Spravodajské
  - Cestovateľské

Na základe uvedeného rozdelenia sa ponúkajú nasledovné otázky:

- Existuje na českom trhu dostatočne veľká skupina ľudí, ktorý majú spoločný záujem, ich záujem nie je dostatočne pokrytý existujúcimi stanicami a majú dostatočnú kúpyschopnosť, aby platili za prémiový obsah?
- Ak nie, ako by mal byť vyskladaný programový obsah všeobecnej televíznej, tak aby dokázal osloviť čo najširšiu masu divákov. Aký je typický český divák ako často sleduje televíziu?

Na uvedené otázky sa pokúsim prostredníctvom dostupných dát a techník data miningu.

Dáta

Dostupné dáta, ktoré budem analyzovať pozostávajú z odpovedí respondentov na **otázky** týkajúce sa ich rôznych životných preferencií, **demografických** údajov o nich a **technickej vybavenosti** ich domácnosti z pohľadu príjmu televízneho signálu. Dataset obsahuje celkovo **2889 pozorovaní** a **51 premenných.** Všetky uvedené premenné sú **kategorické,** okrem premennej počet členov domácnosti.

Premenné, ktoré sú dôležité z pohľadu zadaných cieľov a otázok:

- **Demografické**:
  - **hinc**
    - Čistý príjem domácnosti.
    - Pomôže nám pri určovaní kúpnej sily.
  - **pinc**
    - Čistý príjem jednotlivca.
    - Pomôže nám pri určovaní kúpnej sily.
  - **hint**
    - Frekvencia pripojenia na internet.
    - Internet predstavuje v súčasnosti alternatívu pre tradičné televízne stanice, ale taktiež je to jeden z kanálov, cez ktorý je vysielanie možné šíriť.
  - **agecat**
    - Veková kategória.
    - Je možné, že preferencie divákov sa budú v rôznych vekových kategóriách líšiť.
  - **red**
    - Dosiahnuté vzdelanie.
    - Je možné, že preferencie divákov sa budú v závislé na ich dosiahnutom vzdelaní.
- **Technické vybavenie:**

  Pre novú televíznu stanicu sú tieto informácie dôležité z pohľadu voľby správnej technológie pre šírenie signálu.

  - **heq#1 -** Domácnosť vlastní satelitnú anténu.
  - **heq#9 -** Domácnosť vlastní káblovú televíziu.
  - **q56#14 -** Využíva vlastní káblovú televíziu.
  - **q56#15 -** Využíva satelit.
  - **q56#16 -** Využíva vysielanie v digitálnej kvalite prostredníctvom pevnej telefónnej linky a internetu.
  - **q56#30 -** Využíva set-top-box.

- **Odpovede na otázky**

  27 otázok vo forme výrokov, s ktorými respondent buď súhlasil alebo nesúhlasil.

Možné odpovede:

- Rozhodne súhlasím
- Skôr súhlasím
- Skôr nesúhlasím
- Rozhodne nesúhlasím

Kompletné znenie otázok + ich interpretácia a možné závery o programových preferenciách respondenta, ktoré z nich možno odvodiť:

	Otázka	**Záver**
1.	Zajímají mě teorie a hypotézy	Má rád **dokumentárne** programy
2.	Mám rád(a) šokující lidi a věci	Má rád **reality show**
3.	Mám rád(a) pestrost v mém životě	Má rád **všeobecné stanice** s **pestrým programom**
4.	Rád(a) vyrábím věci, které pak mohu použít každý den	Má rád relácie pre **kutilov**
5.	Přesně jak říká Bible, svět byl skutečně stvořen v šesti dnech	Je **veriaci**
6.	Rád(a) se dozvídám něco o umění, kultuře a historii	Má rád **dokumentárne** programy o **histórii** a **umení**
7.	Často vyhledávám vzrušující zážitky	Má rád programy o **cestovaní** a možno aj o **extrémnych športoch**
8.	Opravdu mě zajímá jen pár věcí	Veľmi **úzko profilovaný** v tom, aký programový obsah sa mu páči
9.	Raději si něco vyrobím, než bych to kupoval(a)	Má rád relácie pre **kutilov**
10.	Mám více schopností než většina lidí	Nezdravo sebavedomý
11.	Považuji se za intelektuála	Má rád **dokumentárne** programy a **spravodajské** relácie
12.	Musím připustit, že se rád(a) předvádím	Má rád **reality show**
13.	Velmi mě zajímá, jak mechanické věci jako např. motory, fungují	Má rád **dokumentárne** programy
14.	Rád(a) se oblékám podle poslední módy	Má rád **fashion** a **lifestyle** programy
15.	Všude okolo nás je dnes příliš sexu	Je konzervatívny
16.	Rád(a) vedu ostatní	Líder
17.	Rád(a) bych strávil(a) rok nebo více v cizí zemi	Má rád programy o **cestovaní**
18.	Rád(a) dělám věci ze dřeva, kovu, či jiného podobného materiálu	Má rád **relácie pre kutilov**
19.	Jsem rád(a) považován(a) za elegantní	Má rád **fashion** programy
20.	Život ženy je naplněn, jen pokud dokáže zajistit šťastný domov pro svou rodinu	?
21.	Mám rád(a) výzvy dělat něco, co jsem ještě nikdy předtím nedělal(a)	Má rád relácie pre **kutilov** a relácie o **cestovaní**

22.	Rád(a) se dozvídám nové věci, i když se mi nemusí k ničemu hodit	Má rád **dokumentárne** programy a **spravodajské** relácie
23.	Rád(a) vyrábím věci vlastníma rukama	Má rád relácie pre **kutilov**
24.	Rád(a) dělám věci, které jsou nové a odlišné	Má rád **netradičné nové relácie a formáty**
25.	Rád(a) si prohlížím hobby prodejny či obchody s auty, autodoplňky	Má rád relácie pre **kutilov** a **auto magazíny**
26.	Rád(a) bych pochopil(a) více o tom, jak funguje vesmír	Má rád **dokumentárne** programy o **vesmíre**
27.	Jsem rád(a), když se můj život týden od týdne příliš neliší	Má rád rutiny a zabehnuté veci

Tabulka č.1

Uvedené závery sú subjektívne odvodené nemusia sa preto zhodovať zo skutočnými preferenciami divákov/respondentov. Toto sú však jediné dáta, z ktorých možno vychádzať pri odvodzovaní toho po akom programovom obsahu je aktuálne dopyt.

- **Ďalšie dáta o sledovaní televízie**
  - **q56**
    - Najviac sledovaná televízia za posledných 7 dní.
  - **q40_35**
    - Ako často sledujete televíziu obecne

Dataset bol je vo veľmi dobrej kvalite – chýbajúce hodnôt buď nie sú žiadne alebo ich chýba menej ako **1%.** Jedinou výnimkou je premenná pinc, v ktorej **chýba 16,82% hodnôt**. Neprítomnosť hodnôt je zrejme spôsobená citlivosťou tohto údaju – osobný príjem. Dataset obsahuje tiež premennú hinc(príjem domácnosti), ktorá nám poskytuje podobnú informáciu o kúpischopnosti a môže v prípade potreby čiastočne nahradiť chýbajúcu hodnotu osobného príjmu. Rozhodol som sa preto chýbajúce hodnoty nejak nedopĺňať a využívať premennú tak ako je.

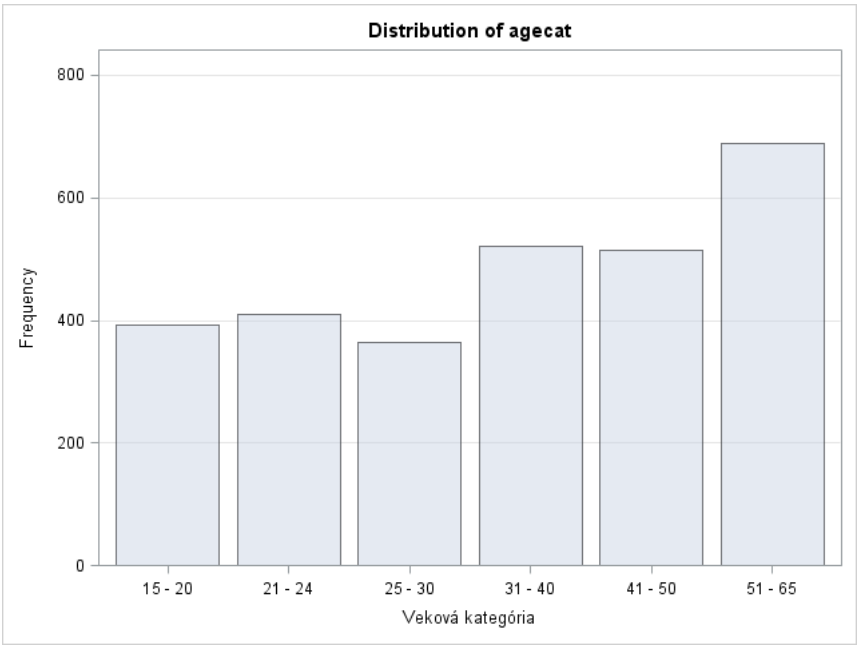
Exploračná analýza

V prvej fáze exploračnej analýzy som sa zameral na demografické charakteristiky respondentov, aby som mal lepšiu predstavu o tom čím je charakteristický priemerný respondent.

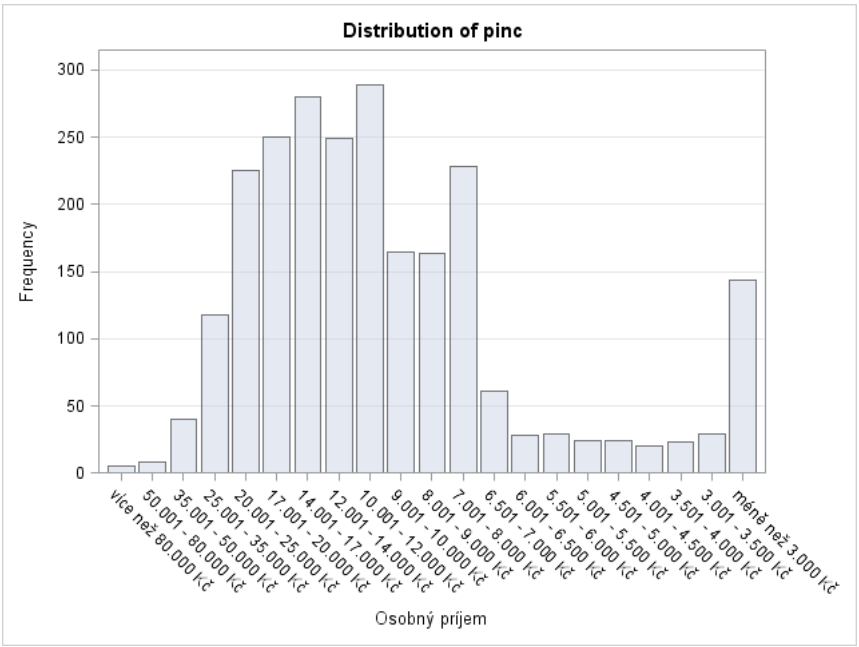
Výsledky:

- Najviac respondentov pochádza z **Moravskoslezkého kraja** – **13,53%**
- Najviac respondentov žije vo **väčších sídlach s 20-100 tisíc** obyvateľmi – **20,73%**
- Najviac respondentov žije v **domácnostiach, ktorých príjem** je **25.001 - 35.000 Kč** – **31%**
- Najviac respondentov žije v domácnostiach, so **4 členmi** – **37,26%**
- Drvivá väčšina respondentov je pripojená na **internet denne** alebo takmer denne – **93,47%**
- Pomer mužov(**47,35%**) a žien(**52,65%**) je medzi respondentmi vyvážený.

- Vekové kategórie:



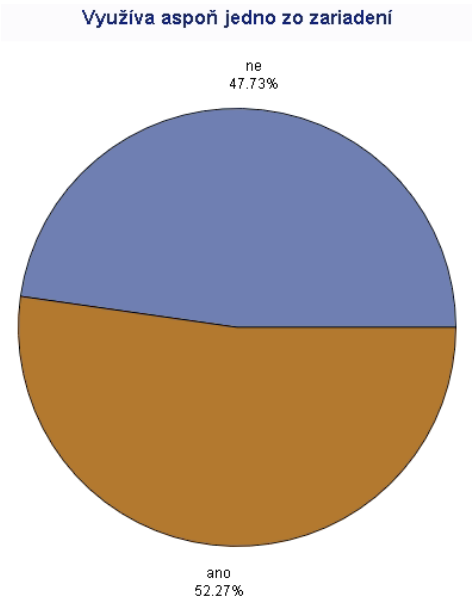
- **24%** respondentov má **vysokoškolské vzdelanie**.
- **50,54%** respondentov je zamestnaných.
- Osobný príjem respondentov



Technické vybavenie domácností z pohľadu príjmu televízneho signálu:

- **27,8%** respondentov využíva **káblovku.**
- **19,66%** respondentov využíva **satelit.**
- **3,67%** respondentov využíva **pevnú linku a internet** na príjem signálu.

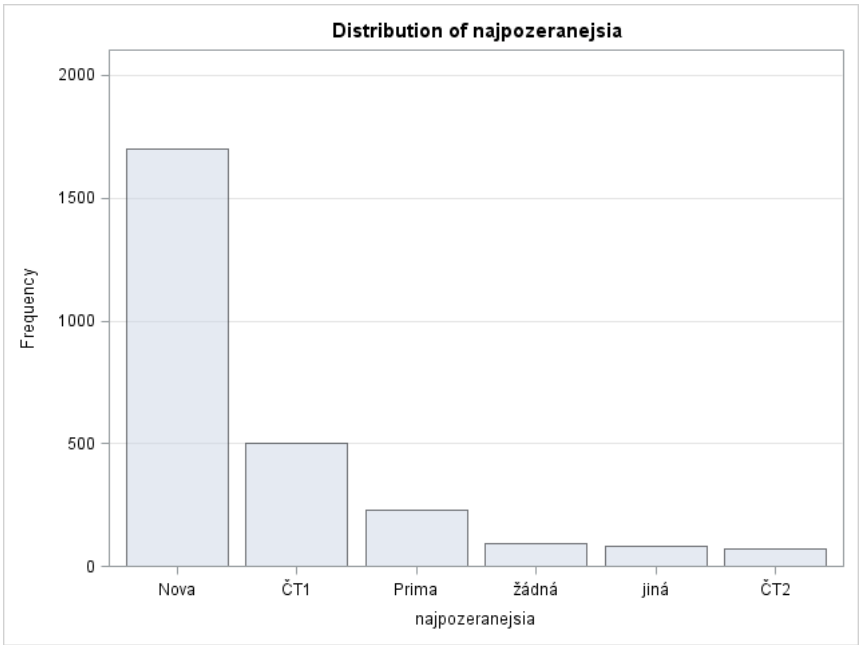
- **10,97%** respondentov využíva **set-top-box.**
- Čo je však podstatné a hlavne pri príjme **úzko profilovaných prémiových staníc**, divák musí **niektorým** z uvedených spôsobov. Takých divákov, ktorí využívajú aspoň jeden z nich je **52,27%.** Toto číslo je ale aj tak relatívne nízke. Môže to byť spôsobené tým, že v datasete nie sú obsiahnuté všetky druhy príjmu signálu.



Typický český divák:

- Televíziu sleduje **denne** alebo takmer denne.
- Za posledných 7 dní pozeral najviac **Novu**.
- Má rád pestrosť v živote – má rád zrejme aj stanice s pestrou škálou relácií a programov.
- Rád sa dozvedá nové veci aj keď sa mu zrovna k ničomu nemusia hodiť.
- Nerád sa predvádza.
- Neverí v biblické stvorenie sveta – neveriaci.

Top 6 najsledovanejších staníc za posledných 7 dní:



- Je zaujímavé ako Nova výrazne prevýšila ostatné stanice. Je to však len sledovanosť za posledných ý dní a nemusí sa zhodovať **s dlhodobou sledovanosťou**. Bolo by ale určite zaujímavé vedieť aké programy vysielala Nova v posledných siedmich dňoch, aby sme mohli skúmať čo tak veľmi divákov oslovilo.
- Podiel stanice ČT2 medzi najsledovanejšími bol **2,42%.** Všetky ostatné stanice mali podiel menej ako **1%.**
- V top 6 sa neumiestnila žiadna úzko profilovaná prémiová stanica, čo je aj logické. Najvyššie bola stanica **HBO** s podielom **0,59%.**

Pridanie nových premenných

Na základe tabuľky č.1 som sa rozhodol do datasetu priradiť nové kategórie respondentov/divákov. Premenné reprezentújuce tieto kategórie majú boolovské hodnoty true/false podľa toho či daný respondent do kategórie patrí alebo nie.

Nové kategórie/premenné:

- **Kutil** – odpovedal na aspoň jednu z otázok **4., 9., 18., 23., 25.** odpoveďou "rozhodne súhlasím".
- **Fashion maniak** – odpovedal na aspoň jednu z otázok **14., 19.**odpoveďou "rozhodne súhlasím".
- **Cestovateľ** – odpovedal na aspoň jednu z otázok **7., 17., 21.** odpoveďou "rozhodne súhlasím".
- **Reality show fan** – odpovedal na aspoň jednu z otázok **2., 12.** odpoveďou "rozhodne súhlasím".
- **Dokumentarista** – odpovedal na aspoň jednu z otázok **1., 6., 11., 13., 22., 26.** odpoveďou "rozhodne súhlasím".

Rozdelenie divákov podľa uvedených kategórií:

Kutil	Fashion Maniak	Cestovateľ	Reality show fan	Dokumentarista
50,57%	24,96%	40,19%	13,12%	64%

- Zaujímavý je najmä výsledok reality show, o ktorých je známe, že sú masovo veľmi populárne. Problém je pravdepodobne v kritériách a nepostačujúcich dátach, ktorými by bolo možné spoľahlivo identifikovať fanúšika reality show.

Čo pozerajú a čo sa zaujímajú diváci s vyšším ako priemerným príjmom

Z pohľadu novej stanice vstupujúcej na trh je toto veľmi **dôležitá skupina divákov**. Sú to diváci, ktorí majú potenciál **platiť za prémiový obsah**, ak by sa rozhodla ísť touto cestou.

Nadpriemerne zarábajúcu skupinu ľudí som definoval nasledovne: Respondent patrí do tejto skupiny, ak má **čistý príjem v horných 25% alebo** ak je členom **domácnosti**, ktorej celkový **čistý príjem je horných 25%.** V tomto prípade to konkrétne znamená, že to sú respondenti, ktorí majú príjem v**yšší ako 17.000 Kč** alebo ich **domácnosť** má **príjem vyšší ako 35.000 Kč**. Taktiež aby bol relevantným možným zákazníkom

Výsledky:

- **Televíziu sledujú denne** alebo takmer denne – **74,84%**
- **Najsledovanejšou** stanicou v tejto skupine bola **tiež Nova**, avšak už s menším podielom než ako tomu bolo v prípade kde boli zahrnutí všetci respondenti.
- Za zmienku stojí tiež to, že Stanica Discovery bola v tomto segmente za posledných 7 dni 9. najsledovanejšia s podielom 1,39%
- Do kategórie "dokumentarista" spadá v tomto segmente **69,28%** respondentov.
- Do kategórie "kutil" spadá v tomto segmente **52,84%** respondentov.
- Ostatné kategórie divákov sú z pohľadu potencionálnej veľkosti trhu málo významné.

Záver

V analýze sa podarilo identifikovať významnú skupinu nadpriemerne zarábajúcich respondentov, ktorých zaujímajú dokumentárne relácie a kutilské relácie. V prípade dokumentárneho obsahu už existuje veľká konkurencia, avšak v prípade kutilov toho programového obsahu až toľko nie je. Toto by mohla byť jedna z ciest, ktorou by sa mohla nová stanica vydať. Bola by však nevyhnutné pokračovať v ďalšej analýze, či je tento segment divákov dostatočne veľký a silný.

Z vytvorenej analýzy ďalej vyplýva, že typický český divák viac inkliminuje k všeobecným staniciam so širokou paletou programov a relácií, preto druhou možnosťou by bola práve cesta pestrého vyváženého programu. Český diváci sledujú televíziu na dennej báze, čo dáva novej stanici stále dobré šance odhryznúť si svoj kúsok z veľkého koláča.

Respondenti uviedli, že na dennej báze používajú tiež internet. V prípade vstupu na trh by mala nová stanici určite zvážiť šírenie svojho vysielania aj cez toto médium.

V ďalšej práci by bolo zaujímavé skúmať preferencie aj ďalších skupín ako sú napríklad študenti, dôchodcovia, nezamestnaní... Bolo by tiež možné skúmať koreláciu medzi odpoveďou na konkrétnu otázku a stanicou, ktorú respondent za uplynulých 7 dní sledoval.