TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

CHURN - Migrace zákazníků ke konkurenci

Semestrální práce

Studijní program: IT – Informační technologie
Studijní obor: AI – Aplikovaná informatika

Autor práce: **David Vancl**
Vedoucí práce:

Liberec 2022

Obsah

1 Úvod

CHURN migrace zákazníků ke konkurenci se týká situace, kdy zákazníci opouštějí danou firmu, produkt nebo službu a přecházejí k jejím konkurentům. Tento jev je důležitým faktorem pro podniky, protože představuje ztrátu tržního podílu a potenciálních příjmů.

Migrace zákazníků ke konkurenci může být způsobena různými faktory, jako je nedostatečná kvalita nebo výkon produktu, nedostatečný zákaznický servis, nepříznivé cenové podmínky, konkurenční nabídka nebo marketingové strategie konkurence, změny preferencí zákazníků nebo inovace na trhu. Je důležité, aby firmy pečlivě sledovaly a analyzovaly migraci svých zákazníků ke konkurenci, aby mohly identifikovat klíčové problémy a přijmout opatření k udržení zákazníků a zlepšení své konkurenceschopnosti.

Efektivní snižování migrace zákazníků ke konkurenci zahrnuje strategie udržení zákazníků, jako je zlepšování kvality produktů a služeb, zvýšení zákaznického servisu, konkurenční cenová politika, vytváření a posilování značky, poskytování výhodných nabídek a slev pro stávající zákazníky, a také aktivní sledování konkurenčního prostředí a rychlá reakce na změny na trhu.

Pochopení důvodů a faktorů, které ovlivňují migraci zákazníků ke konkurenci, je klíčové pro podniky, aby mohly optimalizovat své obchodní strategie a udržet si loajalitu zákazníků.

2 Příprava a instalace prostředí

1. Pro přípravu projektu je využíván IBM SPSS Modeler, instalátor
2. Běhové prostředí, ve kterém poběží program v jazyce R, bude docker container. Je tedy nutné mít nainstalován Docker, instalátor
3. Visual Studio Code je používán pro psaní zdrojových kódů jazyka R, instalátor
4. Do VSC je pořeba doinstalovat balíčky pro práci s dockerem a R jazykem:
   (a) Docker, více zde
   (b) Remote Development, více zde
   (c) R + RDebugger, více zde
5. V konzoli ve složce s projektem otevřu VSCode příkazem: `code .`
6. Uvnitř VSCode otevřu menu pomocí klávesové zkratky `CTRL + SHIFT + P`
   (a) Vyberu `Add Dev Container Configuration Files..`
   (b) Jako image zvolím `r-ver:4.2`
   (c) Výsledný soubor zde
7. Pro spuštění kontejneru a připojení se dovnitř provedu stisknutím klávesové zkratky `CTRL + SHIFT + P` a výběrem `Open Folder in Container...`
8. Po stažení, spuštění a připojení do devcontaineru je prostředí připraveno a nastaveno pro začátek programování v R.

3 Rozbor struktury dat

Vstupní data jsou pro program zadána v textových souborech rawdata a deploydata. Tyto soubory obsahují data oddělená ","a hlavičku popisující jednotlivé sloupce. Soubor `rawdata` je použit pro natrénování jednotlivých modelů a následné otestování. Soubor `deploydata` (oproti `rawdata` nemá sloupec `CHURNED`) je využit k následné predikci modelem C5.0.

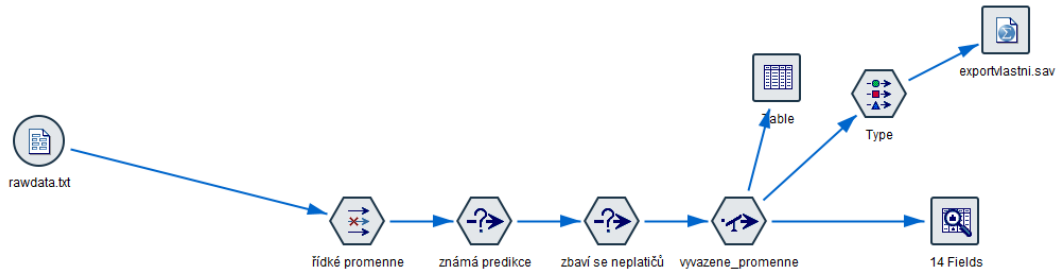
Zde je rozpis, co který sloupec reprezentuje:

- **ID**: Id zákazníka (pro predikci z hlediska modelů je nepotřebný)
- **LONGDIST**: Počet provolaných minut v meziměstských hovorech
- **International**: Počet provolaných minut v mezinárodních hovorech
- **LOCAL**: Počet provolaných minut v lokálních hovorech
- **DROPPED**: Počet přerušených hovorů
- **PAY_MTHD**: Způsob platby
- **LocalBillType**: Tarif pro místní volání
- **LongDistanceBillType**: Tarif pro meziměstské volání
- **AGE**: Věk uživatele
- **SEX**: Pohlaví uživatele
- **STATUS**: Stav uživatele
- **CHILDREN**: Počet dětí uživatele
- **Est_Income**: Příjem uživatele
- **Car_Owner**: Zda uživatel vlastní auto
- **No phone lines**: Žádné telefonní číslo?
- **CHURNED**: Identifikátor odchodu ke konkurenci (Current - stálý, Vol - přelétavý, InVol - neplatič)

4 Vizualizace v Modeleru

Celý projekt je rozdělen do tří částí, které byly připraveny v prostředí SPSS Modeleru a sestaveny. Jednotlivé streamy jsou k dispozici ve složce `Models`. Tyto části odpovídají struktuře kódu v jazyce R.

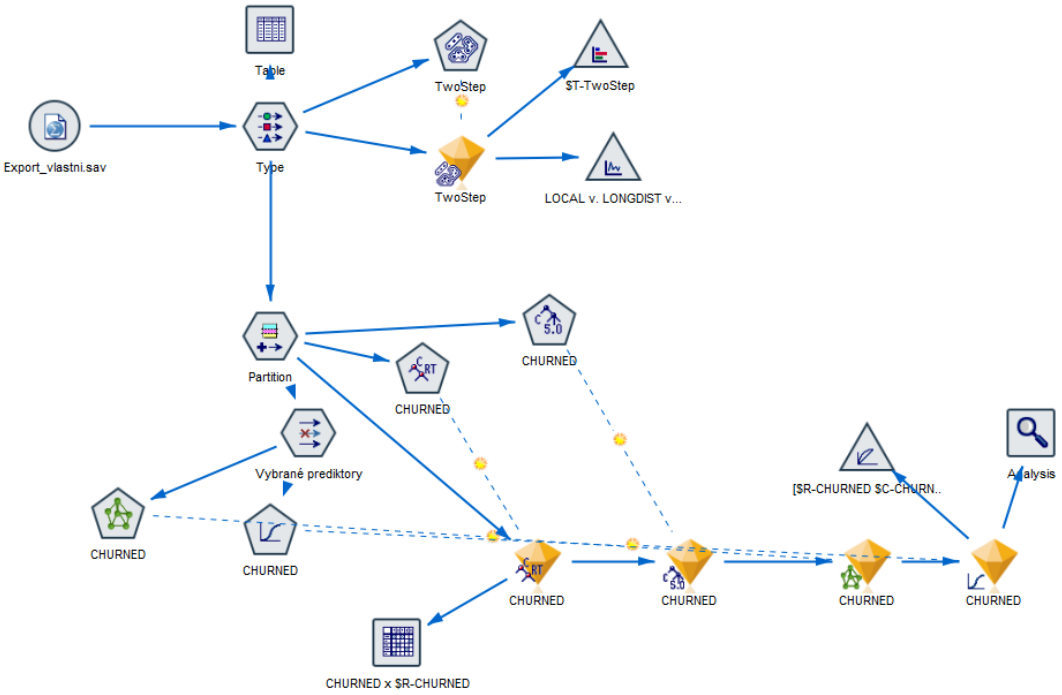
1. Celý projekt začíná přípravou dat, která je vymodelována v souboru `Preparation.str`.



Obrázek 4.1: Model přípravy dat

   (a) Data jsou v prvním uzlu načtena z textového souboru `rawdata.txt`.
   (b) Druhý uzel odebírá nepotřebný sloupec `No.phone.lines`.
   (c) Třetí uzel provede filtrování dat podle sloupce `CHURNED`, kde jsou prázdné hodnoty.
   (d) Čtvrtý uzel odebírá řádky "neplatičů".
   (e) Pátý uzel provede vyvážení dat podle hodnot ve sloupci `CHURNED` (duplikuje některé řádky).
   (f) Tabulka "šmírovačka"zobrazuje data po aktuálním uzlu.
   (g) Šestý uzel nastavuje sloupce `ID` jako nepodstatné a `CHURNED` jako cíl pro hledání.
   (h) Poslední uzel přebírá připravená data a ukládá je do souboru`Export_vlastni.sav`.

2. Druhý stream `Modeling.str` vizualizuje natrénování čtyř různých modelů na 90



Obrázek 4.2: Modelování dat

(a) Připravená data z předchozího streamu jsou načtena.

(b) Uzel `Type` specifikuje sloupce `ID` jako nepodstatné a `CHURNED` jako cíl.

(c) Data jsou přiřazena do jednotlivých clusterů pomocí algoritmu `TwoStep` na základě parametrů `LONGDIST`, `International` a `LOCAL`.

(d) Následují grafy a analýza na základě algoritmu `TwoStep`.

(e) Uzel `Partition` rozděluje data na 90

(f) Model C5.0 je trénován na datech.

(g) Model C&R Tree je trénován na datech.

(h) Pro další dva modely, které pracují pouze s numerickými hodnotami, jsou data převedena na číselné hodnoty nebo jsou odebrána.

(i) Model logistické regrese je trénován na datech.

(j) Model neuronové sítě je trénován na datech.

(k) Vygenerované natrénované modely jsou sériově spojeny.

(l) Poslední uzly obsahují analýzu testování na jednotlivých modelech a jejich graf.

3. Třetí a poslední stream `Deployment.str` aplikuje model s algoritmem C5.0 na datech a vyhodnocuje jednotlivé zákazníky pomocí spočítaného skóre.

(a) Nejdříve jsou načtena zadaná data ze souboru `deploydata.txt`.

Results for output field CHURNED

Individual Models

Comparing $R-CHURNED with CHURNED

'Partition'	1_Training		2_Testing	
Correct	1 340	89,33%	147	89,09%
Wrong	160	10,67%	18	10,91%
Total	1 500		165	

Comparing $C-CHURNED with CHURNED

'Partition'	1_Training		2_Testing	
Correct	1 355	90,33%	148	89,7%
Wrong	145	9,67%	17	10,3%
Total	1 500		165	

Comparing $N-CHURNED with CHURNED

'Partition'	1_Training		2_Testing	
Correct	1 232	82,13%	135	81,82%
Wrong	268	17,87%	30	18,18%
Total	1 500		165	

Comparing $L-CHURNED with CHURNED

'Partition'	1_Training		2_Testing	
Correct	1 158	77,2%	124	75,15%
Wrong	342	22,8%	41	24,85%
Total	1 500		165	

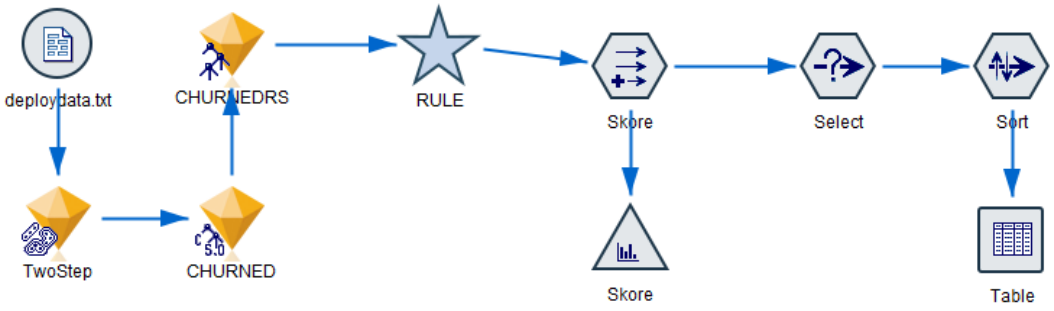
Agreement between $R-CHURNED $C-CHURNED $N-CHURNED $L-CHURNED

'Partition'	1_Training		2_Testing	
Agree	1 242	82,8%	130	78,79%
Disagree	258	17,2%	35	21,21%
Total	1 500		165	

Comparing Agreement with CHURNED

'Partition'	1_Training		2_Testing	
Correct	1 116	89,86%	118	90,77%
Wrong	126	10,14%	12	9,23%
Total	1 242		130	

Obrázek 4.3: Analýza jednotlivých modelů



Obrázek 4.4: Model nasazení

(b) Je vložen vygenerovaný model pro clusterizaci `TwoStep`, který přidává sloupec s přiřazeným clusterem.

(c) Následuje natrénovaný model s algoritmem C5.0, který přidává predikovanou hodnotu a její pravděpodobnost.

(d) Uzel `Skore` přidává sloupec s hodnotou vypočítanou z predikované hodnoty a pravděpodobnosti. Skóre udává "spolehlivost"zákazníka.

(e) Uzel `Select` filtruje zákazníky jejichž skóre je větší než 0.5.

(f) Uzel `Sort` řadí záznamy podle clusteru a skóre.

(g) Nakonec jsou data vypsána do tabulky.

5 Implementace v jazyce R

Jednotlivé zdrojové kódy jsou umístěny ve složce `R`. V kořenu této složky jsou umístěny 3 hlavní skripty, které odpovídají výše uvedeným 3 streamům: příprava dat, modelování a vyhodnocení na deploymenových datech.

V adresáři `Utils` je k dispozici soubor `install.R`, který slouží k instalaci potřebných balíčků pro správný běh programu. Stačí tento skript spustit jednou po vytvoření kontejneru. Balíčky jsou instalovány lokálně pro daný kontejner.

Každý hlavní skript má v hlavičce nastavenou základní cestu, odkazy na potřebné soubory a import utility, která se nachází v adresáři `Utils`. Tato utility obsahuje dodatečné funkce pro manipulaci s daty.

Hlavní tělo programu je přehledně rozděleno do jednotlivých kroků, které odpovídají popsaným částem v Modeleru. V samotném kódu jsou příslušné řádky komentovány. Níže se zaměříme především na dosažené výsledky implementace.

Preparation.R stream

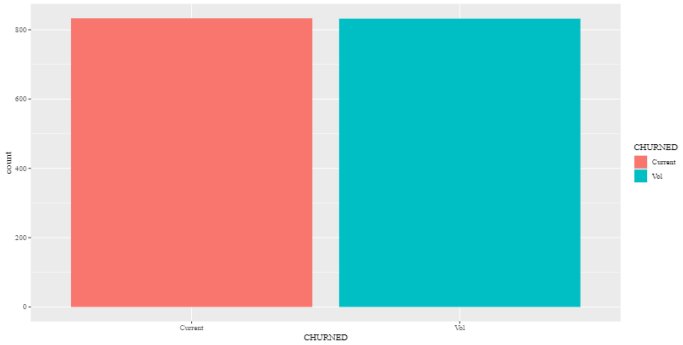
Implementace postupu je uvedena v číslovaném seznamu níže:

1. Načtení dat (DATA) a zobrazení náhledu
2. Filtrace sloupce No.phone.lines a zobrazení náhledu
3. Výběr pouze neprázdných hodnot v sloupci CHURN a zobrazení náhledu
4. Vyloučení neplatících zákazníků na základě sloupce CHURN
5. Vyvážení dat pro sloupec CHURN:
   (a) Určení maximálního počtu výskytů mezi hodnotami
   (b) Inicializace prázdné tabulky pro vyvážená data
   (c) Pro každou hodnotu ve sloupci:
      i. Získání počtu výskytů této hodnoty
      ii. Výpočet počtu potřebných nových řádků
      iii. Výběr náhodných řádků s touto hodnotou pro duplikaci
      iv. Přidání nových řádků do vyvážené tabulky
   (d) Sloučení původní tabulky s vyváženou tabulkou
   (e) Seřazení podle sloupce ID

(f) Zobrazení grafu vyváženosti a tabulky s daty

6. Převedení faktorů na řetězce a uložení výsledku do souboru `export_data_r.sav`

Výsledkem je dataframe s vyváženými daty pro sloupec CHURNED. Graf vyváženosti dat je zobrazen níže:

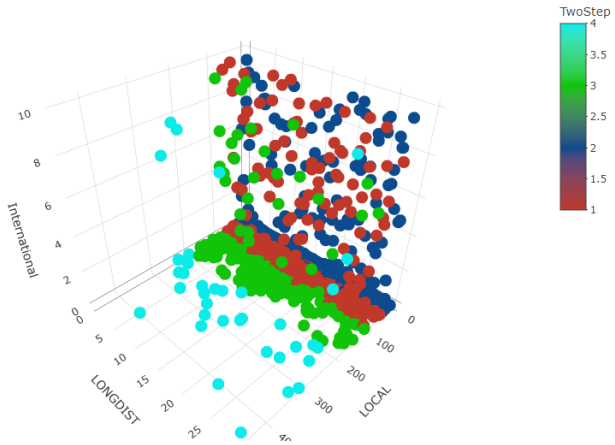


Obrázek 5.1: Graf vyváženosti dat

Modeling.R stream

Implementace postupu je uvedena v číslovaném seznamu níže:

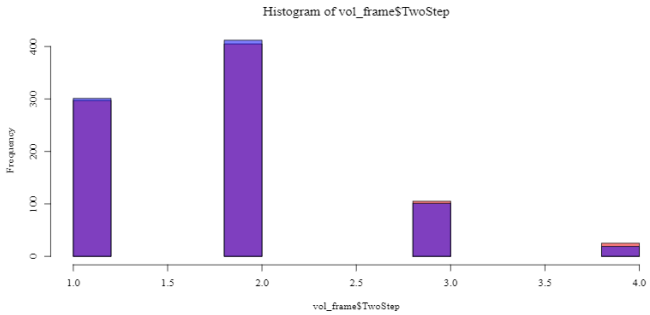
1. Načtení vyexportovaných dat
2. Odebrání sloupce s irelevantním ID
3. Vytvoření rozdělení pomocí clusterů (alternativa TwoStep)
4. Vykreslení 3D grafu pro zařazení do clusterů



Obrázek 5.2: Graf clusterovaných dat

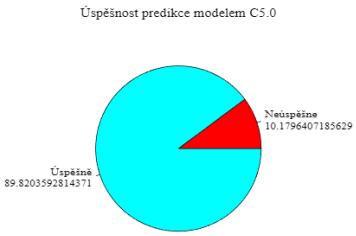
5. Vykreslení histogramu pro clusterované hodnoty

(a) Hodnoty vztažené k "Vol"(přelétaví zákazníci)
(b) Hodnoty vztažené k "Current"(stálí zákazníci)



Obrázek 5.3: Histogram hodnot

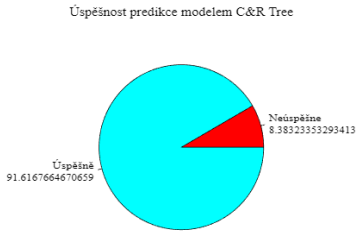
6. Rozdělení dat na trénovací a testovací
   (a) Náhodné zamíchání řádků v dataframe
   (b) Rozdělení na 90% trénovací data a 10% testovací data
   (c) Vytvoření kopie testovací tabulky
7. Predikce pomocí algoritmu C5.0
   (a) Příprava trénovacích dat a oddělení cílového sloupce
   (b) Natrénování modelu
   (c) Příprava testovacích dat a oddělení cílového sloupce
   (d) Predikce na testovacích datech
   (e) Vyhodnocení výsledků pomocí koláčového grafu



Obrázek 5.4: Vyhodnocení výsledků

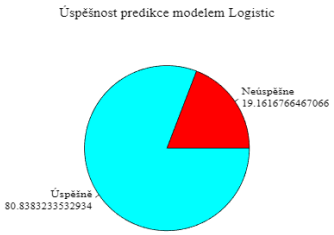
8. Predikce pomocí algoritmu C&R Tree
   (a) Natrénování modelu

(b) Predikce na testovacích datech
(c) Uložení výsledků do originální tabulky
(d) Výpis výsledků predikce



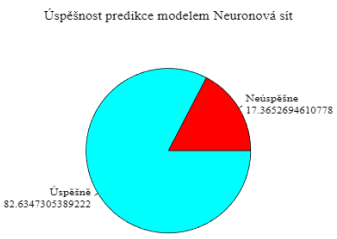
Obrázek 5.5: Vyhodnocení výsledků

9. Predikce pomocí algoritmu Logistic
   (a) Převod na numerická data
   (b) Rozložení testovacích dat a sloupců s výsledkem
   (c) Natrénování modelu
   (d) Predikce na testovacích datech a převod výsledků na 0 a 1
   (e) Uložení výsledků do originální tabulky a převod na text
   (f) Výpis výsledků predikce



Obrázek 5.6: Vyhodnocení výsledků

10. Predikce pomocí algoritmu Neuronová síť
    (a) Natrénování modelu
    (b) Predikce na testovacích datech a převod výsledků na 0 a 1
    (c) Uložení výsledků do originální tabulky a převod na text
    (d) Výpis výsledků predikce
11. Výpis celkové tabulky a predikovaných hodnot

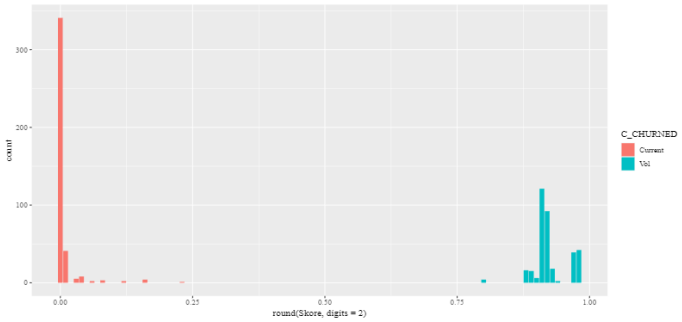


Obrázek 5.7: Vyhodnocení výsledků

Deployment.R stream

Implementace postupu je uvedena v číslovaném seznamu níže:

1. Načtení dat
2. Odstranění sloupců ID a No.phone.lines
3. Vytvoření rozdělení pomocí clusterů (alternativa TwoStep)
4. Provedení predikce na základě modelu C5.0
5. Přidání procentuální hodnoty k datům
6. Přidání vypočítaného skóre do tabulky
7. Vykreslení grafu pro rozložení skóre



Obrázek 5.8: Vyhodnocení výsledků

8. Filtrace řádků s skóre vyšším než 0.5
9. Doplnění sloupce ID zpět do dat pro identifikaci
10. Seřazení dat vzestupně podle clusteru a sestupně podle skóre
11. Výpis vyhodnocených dat

6 Závěr

Migrace zákazníků ke konkurenci byl zaměřen na analýzu a predikci odchodu zákazníků a jejich migrace ke konkurenčním společnostem. Cílem projektu bylo identifikovat klíčové faktory a chování zákazníků, které přispívají k odchodu, a vytvořit modely pro predikci budoucího chování zákazníků.

V rámci projektu byly použity různé algoritmy a techniky analýzy dat, včetně TwoStep clustering, C5.0, C&R Tree a Logistic regression. Tyto algoritmy umožnily identifikovat shluky zákazníků, vyhodnotit jejich charakteristiky a predikovat pravděpodobnost odchodu zákazníků ke konkurenci.

Během projektu byly také provedeny vizualizace dat, které umožnily lepší porozumění distribuci a vztahům mezi proměnnými. Dále byly vyhodnoceny výsledky predikcí a úspěšnost modelů. Výsledkem projektu byl také výpočet skóre pro jednotlivé zákazníky, které odráželo jejich pravděpodobnost migrace ke konkurenci.