**УНИВЕРСИТЕТ ЗА НАЦИОНАЛНО И СВЕТОВНО СТОПАНСТВО**

**Факултет „Приложна информатика и статистика“**

**Катедра „Информационни технологии и комуникации“**

# АНАЛИТИЧЕН МОДЕЛ ЗА ОЦЕНКА НА РИСКА ОТ ОТПАДАНЕ НА КЛИЕНТИ

## АВТОРЕФЕРАТ

на дисертационен труд за присъждане
на образователна и научна степен „доктор“

Докторант: **Христо Димитров Янчев**

Научен ръководител: **доц. д-р Дорина Кабакчиева**

София, 2026

Дисертационният труд се състои от увод, три глави, заключение, приноси, списък с таблици, списък с фигури и библиография. Той е в общ обем от 157 страници. В изложението присъстват 24 таблици и 69 фигури.

Дисертационният труд е обсъден в катедра „Информационни технологии и комуникации" при Университет за национално и световно стопанство – София и е насочен за защита пред научно жури. Авторът на дисертационния труд е задочен докторант в катедрата.

Защитата на дисертационния труд ще се състои на 27.03.2026 г. от 10:00 часа в зала 2032А на Университет за национално и световно стопанство – София на заседание на научното жури. Материалите по защитата са на разположение на интересуващите се в Университет за национално и световно стопанство – София.

# Съдържание

# Обща характеристика на дисертационния труд

## 1. Обем и структура на дисертационния труд

Дисертационният труд е структуриран в три глави, както следва:

В **първа глава** е направен обзор на актуалното състояние на Data Mining областта. Разгледани са типовете Data Mining задачи, като фокусът е върху задачата за класификация, която е подходяща за решаването на бизнес проблема с напускащите клиенти. Избрани и представени са три Data Mining метода за решаване на задачата за класификация при откриване на скрити знания в данни от сферата на телекомуникациите. Описани са различни методи за оценка и сравнение на получените класификатори. Разгледани са съществуващи подходи за реализация на Data Mining проекти. Изброени са съществуващи софтуерни решения за реализирането на подобни проекти. Дефинирана е целта на научното изследване.

Във **втора глава** е представена методиката за изпълнението на научното изследване. Обявен е избраният подход за изпълнението на Data Mining проекта. Изброени са софтуерите, подходящи за реализирането на поставените задачи, и са описани подробно данните, които участват при решаването на задачата за класификация. Изброени са извършените дейности по предварителната подготовка на данните за клиенти на компания от сферата на телекомуникациите.

В **трета глава** е представено изграждането на Data Mining моделите в RapidMiner Studio, избрания софтуерен инструмент за реализация на проекта. Приложени са вече избраните Data Mining методи върху потребителските данни за клиенти на телекомуникационна компания. Оптимизирани са представените модели за класификация и са оценени получените резултати, използвайки избраните метрики за оценка.

Дисертационният труд се състои от увод, три глави, заключение, приноси, списък с таблици, списък с фигури и библиография. Той е в общ обем от 157 страници. В изложението присъстват 24 таблици и 69 фигури.

## 2. Актуалност на темата на изследвания проблем

Напускащи клиенти (Customer Churn) е термин, използван за определяне на клиенти, които приключват договорите или абонаментите си с дадена компания. Това е един от най-големите проблеми в различни индустрии, свързани с финансите, търговията на дребно, пътешествията, телекомуникациите и др. Необходимо е да се разбере какво причинява този проблем, за да може една компания да има непрекъснат и устойчив растеж на бизнеса. Определянето на тези клиенти е трудна задача, затова всички налични клиентски данни в една компания трябва да бъдат анализирани. Удовлетвореността на клиентите, нивото на ангажираност на клиентите или намаляването на използването на услугите на компанията може да са знаци за потенциален такъв проблем. В тази разработка са разгледани различни причини клиентите да напуснат компанията, към която имат абонаменти.

Според Хаден (Hadden, J., Tiwari, A., Roy, R., Ruta, D., 2007), поради високите разходи, целта на телекомуникационните компании се е променила от придобиване на нови клиенти към задържане на съществуващи. Това означава, че спирането на напускащите клиенти би довело до увеличаване на продажбите и намаляване на пазарните разходи, в сравнение с привличането на нови клиенти. Затова прогнозата за напускащи потребители става все по-популярна сред бързо развиващите се и конкурентни компании в телекомуникационния сектор. Това прави моделите за прогнозиране на напускащи клиенти неразделна част от процесите на вземане на решения и планиране в сферата.

През 2000 г. разходите, определени за привличане на нови клиенти, са били три до пет пъти по-високи от разходите за задържане на съществуващи клиенти (Ng, K., & Liu, H., 2000). През 2007 г. разходите за привличане на нови клиенти вече са били пет до десет пъти по-големи от разходите за запазване на настоящите такива (Chu, B.H., Tsai, M.S., Ho, C.S., 2007). Поради това, задържането на клиентите се превръща в толкова важна тема за множество бизнеси през годините.

Разходите за привличане на нови клиенти се увеличават всеки ден. Така в маркетинговото подразделение на телекомуникационните компании започна да расте нова идея. Кампаниите за печелене на нови клиенти вече не са с най-висок приоритет, за разлика от кампаниите, създадени за развиване на силни взаимоотношения клиент-компания. В момента компаниите са фокусирани върху програми за удовлетвореност на клиентите и програми за лоялност на клиентите.

## 3. Обект и предмет на дисертационния труд

**Обект** на настоящата дисертация са напускащите клиенти в сферата на телекомуникациите и по-конкретно техните потребителски данни, върху които да бъда приложени Data Mining методи за класификация.

**Предмет** на този дисертационен труд е разработването на методология за откриването на потенциалните напускащи клиенти на телеком компании, разработвайки и оптимизирайки Data Mining модели за класификация с цел навременното откриване на тези клиенти.

Тази дисертация цели да покаже приложимостта и резултатността на различните Data Mining модели за класификация и тяхното реално приложение. Терминът Data Mining се използва вместо българския превод – извличане на скрити знания в данни, с цел по-кратък и последователен изказ, в текста на дисертационния труд. Когато се споменават Data Mining модели за класификация, това са обучени класификатори, които се получават чрез прилагане на методи за класификация при откриването на скрити знания в големи обеми от данни.

## 4. Цел, задачи и работна хипотеза

**Целта** на настоящата дисертация е да се генерират, изследват и подобрят Data Mining модели за класификация, като се разработи аналитичен модел за откриване на клиенти, прекратяващи своя договор (напускащи клиенти), на базата на потребителски данни от сферата на телекомуникациите, включващи демографски характеристики, данни за използваните от тях услуги, типа и метода на плащане, месечните и общите им плащания.

За постигане на посочената цел се поставят следните **задачи**:

- Разработване на методика за решаване на Data Mining задача за класификация;
- Избор на подходящи софтуерни инструменти за реализация на Data Mining задача за обучение на класификатори;
- Генериране на модели за извличане на скрити знания от данни, използвайки подходящи методи за решаване на Data Mining задачата за класификация;
- Оценяване на генерираните класификатори, използвайки избрани метрики за оценка;
- Подобряване на точността на предсказване на получените класификатори;
- Сравняване на получените Data Mining модели за класификация и препоръка за използване при решаване на бизнес проблема за откриване на напускащи клиенти в сферата на телекомуникациите.

Реализирането на това изследване в дисертацията цели да покаже детайлни стъпки по реализацията на Data Mining проект и да докаже, че методите и инструментите могат да бъдат прилагани независимо от сферата.

Представената цел на дисертацията дава възможност за формулиране на следната **хипотеза**: Възможно е обучените класификатори да бъдат успешно използвани за предсказване класа на отпадащите клиенти на телекомуникационна компания.

## 5. Приноси на дисертационния труд

***Научни приноси***

Най-важните *научни приноси*, с характер *обогатяване на съществуващи знания,* включват:

- Предложена е методика за решаване на бизнес проблем, свързан с навременното откриване на потенциалните напускащи клиенти в сферата на телекомуникациите, чрез решаване на Data Mining задачата за обучаване на класификатори;
- На базата на подходящо избраната методика, включваща CRISP-DM подход и софтуерен инструмент RapidMiner Studio, са обучени, оценени и подобрени класификатори чрез използване на три различни метода - „Дърво на решенията”, „Невронни мрежи“ и „Логистична регресия“, като най-висока точност на

предсказване на класа на отпадащите клиенти се получава за класификатора, получен чрез метода „Дърво на решенията”;

- Получени са потвърдителни факти за изследваната съвкупност от данни, като точността на предсказване на получените модели за класификация на отпадащи клиенти в сферата на телекомуникациите е сравнима с резултати, получени от други изследователи, използвали подобни данни и Data Mining методи за класификация (Табл.1.4).

***Научно-приложни приноси***

Най-важните *научно-приложни приноси*, с характер *приложение на научните постижения в практиката,* включват:

- Получени са множество резултати от откриване на знания в данни чрез обучение на класификатори, за откриване на отпадащи клиенти на телекомуникационни компании, описващи зависимости между вероятността за отпадане на клиентите и техни характеристики като типа на договора (Contract), продължителността на договорите на клиентите (tenure) и платените от тях сметки (MonthlyCharges, TotalCharges), типа на интернет услугата (Internet Service);
- Апробирана е методика на изследването, включваща предварителна подготовка на данните, обучение на класификатори за откриване на знания в данни, и оценка и сравнение на получените модели, която може да се използва и в други случаи, за извличане на знания от данни в сферата на телекомуникациите;
- Използваният подход и методика на изследвания, може да бъде приложена върху нови различни данни, което да доведе до успешно идентифициране на отпадащи клиенти не само в сферата на телекомуникациите, но и в други области като застраховане, облачни и хостинг услуги, стрийминг услуги като Netflix, Disney+, HBO и други, при които съществува подобен модел за месечен или годишен абонамент.

# Кратко изложение на дисертационния труд

# 1. Първа Глава: Обзор и актуално състояние на областта извличане на скрити знания от данни (Data Mining) и методи за класификация

## 1.1. Актуалност и проблем на научното изследване

Този дисертационен труд е посветен на решението на проблема *Напускащи Клиенти*, в сферата на телекомуникациите. Това са потребители, които прекратяват договора си към дадена телеком копания. Този проблем е от голямо икономическо значение за фирмите в тази сфера, защото откриването и придобиването на нови клиенти е в пъти по-скъпо от задържането на вече съществуващите такива. Откриването на тези клиенти не е лесна задача и е необходим специфичен подход, за да бъдат дефинирани правилно стъпки, през които изследователите да преминат, за да получат отговор на този въпрос. Този проблем обикновено се решава с помощта на Data Mining задачи.

В статията „Data Mining Applications in Customer Churn Management“ (KhakAbi, S., Gholamian, M.R., & Namvar, M., 2010) е представено проучване относно различните методи, използвани за анализ на напускащите клиенти. На базата на анализираните избрани научни публикации по темата, се стига до следното заключение относно честотата на използване на различните методи за класификация при изграждане на модел за напускащите клиенти (Табл.1):

*Таблица 1. Петте най-използвани методи за „Data Mining“ според статията „Data Mining Applications in Customer Churn Management“*

| Методи | Брой публикации |
|---|---|
| Neural Networks | 15 |
| Decision Tree | 13 |
| Logistic Regression | 13 |
| Random Forests | 7 |
| Support Vector Machine | 7 |

Най-често ползваният метод е „Невронни мрежи“, поради високата успеваемост на моделите. Следващ по ред е методът Дърво на решенията, заради възможността получените резултати да бъдат лесно обвързани с критериите за тяхното постигане. Следващите методи от класацията биват избирани по-рядко от научните изследователи, заради по-ниската точност на моделите и по-трудното им тълкуване в последствие.

Представените методи се прилагат към данни от много различни сектори. Основният фокус на този дисертационен труд обаче, е върху областта на телекомуникациите. От публикациите, показани в Таблица 1.1, са селектирани само такива, които са свързани със сферата на телекомуникациите. Отново методите са подредени според броя публикации, в които се срещат (Таблица 2):

*Таблица 2. Брой публикации с данни от сферата на телекомуникациите, подредени според използвания Data Mining метод*

| **Методи** | **Брой публикации** |
|---|---|
| Neural Networks | 5 |
| Decision Tree | 3 |
| Logistic Regression | 2 |
| Random Forests | 1 |
| Support Vector Machine | 1 |

От втората сравнителна таблица, която е подредена по брой публикации, свързани с данни от сферата на телекомуникациите, е видимо, че класирането на методите не се е променило, поради вече споменатите характеристики.

Допълнително са анализирани и други публикации, които не са част от горните сравнения, по темата за напускащи клиенти на телеком компании, като получените резултати от анализа са обобщени в Таблица 3:

*Таблица 3. Допълнително анализирани публикации, използващи Data mining методи при откриване на скрити знания в данни от сферата на телекомуникациите*

| **Методи** | **Автори** | **Брой публикации** |
|---|---|---|
| *Neural Network (and Hybrid Neural Network* | (Mozer, Wolniewicz, Grimes, & Johnson, 2000), (Tsai, C.F., Lu, Y.H., 2009), (Mamcenko, J., & Gasimov, J., 2014), (Vuval, 2020), (Khalid, Abdulazeez, Zeebaree, Ahmed, & Asaad, 2021), (Boujelbene, Labidi, Lemjid, Ncib, & Zghal, 2024) | 6 |
| *Decision Tree* | (Tuğba, U., Gürsoy, Ş., 2010), (Owczarczuk, 2010), (Khalid, Abdulazeez, Zeebaree, Ahmed, & Asaad, 2021), (Saleh & Abd-Alsabour, 2024), (Boujelbene, Labidi, Lemjid, Ncib, & Zghal, 2024), (Saini & Garg, 2017) | 6 |
| *Logistic regression* | (Mozer, Wolniewicz, Grimes, & Johnson, 2000), (Tuğba, U., Gürsoy, Ş., 2010), (Khan, M. R., | 5 |

| | | |
|---|---|---|
| | Manoj, J., Singh, A., & Blumenstock, J., 2015), (Owczarczuk, 2010), (Boujelbene, Labidi, Lemjid, Ncib, & Zghal, 2024) | |
| *Random Forests* | (Khan, M. R., Manoj, J., Singh, A., & Blumenstock, J., 2015), (Khalid, Abdulazeez, Zeebaree, Ahmed, & Asaad, 2021), (Saleh & Abd-Alsabour, 2024), (Boujelbene, Labidi, Lemjid, Ncib, & Zghal, 2024) | 4 |
| *Support Vector Machines* | (Khan, M. R., Manoj, J., Singh, A., & Blumenstock, J., 2015), (Boujelbene, Labidi, Lemjid, Ncib, & Zghal, 2024) | 2 |
| *K-Nearest Neighbors* | (Khan, M. R., Manoj, J., Singh, A., & Blumenstock, J., 2015), (Boujelbene, Labidi, Lemjid, Ncib, & Zghal, 2024) | 2 |

Данните в Таблица 3 показват, че тенденцията да бъдат използвани „Невронни мрежи“, „Дърво на решенията“ и „Логистична регресия“ се запазва, и това са най-използваните методи за създаване на класификатори при работа с данни от сферата на телекомуникациите.

Направен е допълнителен обзор на резултатите от публикации, в които се използва съвкупността от данни (Kaggle, 2016), която е изследвана и в този дисертационен труд. Целта е потвърждение и сравнение на общата точност и точността на предсказване на класа на напускащите клиенти, представени в тези научни публикации, спрямо постигнатите точности на предсказване в тази дисертация (Таблица 4).

*Таблица 4 Представени са резултатите на точността на предсказване на двата класа и точността на предсказване на класа на напускащите клиенти от публикации, в който е използвана изследваната съвкупност от данни*

| **Име на публикацията** | **Използвани методи** | **Точност на предсказване на двата класа** | **Точност на предсказване на класа на отпадащите клиенти** |
|---|---|---|---|
| *Telco Customer Churn Prediction (Wei, 2024)* | Random Forest, SVM, Logistic regression | Random Forest: **81.17%**<br>SVM: **81.95%**<br>Logistic regression: **81.7%** | Random Forest: -<br>SVM: -<br>Logistic regression: - |
| *Churn prediction analysis of telecom customers using svm, random forest and logistic regression* | Random Forest, SVM, Logistic regression | Random Forest: **76%**<br>SVM: **74%**<br>Logistic regression: **79%** | Random Forest: **81%**<br>SVM: **83%**<br>Logistic regression: **83%** |

| | | | |
|---|---|---|---|
| *models using orange data mining tools (Nurtriana, 2024)* | | | |
| *Explaining customer churn prediction in telecom industry using tabular machine learning models (Poudel, Pokharel, & Timilsina, 2024)* | Neural Networks, Random Forest, Logistic regression | Neural Networks: **74% ± 6%**<br>Logistic regression: **79% ± 2%**<br>Random Forest: **80% ± 2%** | Neural Networks: **58% ± 26%**<br>Logistic regression: **64% ± 4%**<br>Random Forest: **71% ± 4%** |

**Анализ на напускащи клиенти**

В областта на телекомуникациите напускащ клиент е някой, който преминава от своя доставчик на услуги към друг, конкурентен доставчик. Задържането на този вид клиенти е основен проблем, а също и голямо предизвикателство на силно конкурентните пазари. Въпреки че напускащите клиентите са хора, които решават да изберат друг доставчик на услуги, вероятно при по-добри условия, тези клиенти не са идентични. Те се класифицират като три типа (Lazarov, V., & Capota, M., 2007):

- *Активни* (доброволни) - това са клиенти, които решават сами да прекратят договора;
- *Пасивни* (недоброволни) - договорът се анулира от самата компания;
- *Циркулиращи* (случайни) - в този случай клиентът няма намерение да премине към друг доставчик.

Необходимо е да се разбере защо клиентът напуска настоящия си телекомуникационен доставчик и отива при друг, затова се извършва анализ на поведението на напускащите клиенти. Анализът включва създаване на шаблон или модел на действията на клиентите. Проблемът е ключов, тъй като е свързан със загуба на приходи и големи разходи за привличане на нови клиенти.

Напускащите клиентите са един от най-големите проблеми в телекомуникационната индустрия, затова всички компании в този сектор предлагат възможно най-много минути и мегабайта в своите планове на възможно най-ниска цена, за да провокират клиентите да сменят оператора (Weiss, 2009).

Откриването на потенциалните напускащи клиенти се извършва чрез прилагане на методите за откриване на знания в данни (Data Mining) чрез обучение на класификатори, като процесът включва няколко основни стъпки:

1. Първата стъпка, която трябва да се направи, е изследването на поведението на клиентите, за да се установят техните нужди;

2. Втора стъпка започва със събирането на данни от изходните системи;
3. Третата стъпка е свързана с анализ на поведението на клиентите;
4. Четвъртата стъпка е предварителен анализ на данните;
5. Последната стъпка е, когато тези данни са налични да се заредят в базата данни на компанията. Накрая е важно да се интерпретират и оценят получените резултати.

**Източници на данни**

За да могат да бъдат анализирани данните на напускащите потребители в сферата на телекомуникациите, трябва първо да бъдат намерени подходящи източници на тези данни. Източници на данни са:

- Социални мрежи
- Мобилни устройства
- RFID (Radio Frequency Identification - Радиочестотна идентификация)
- Сателити
- Дигитални сензори
- Интернет на обектите (Internet-of-Things – IoT)
- Интернет сайтове
- Физически магазини
- e-Commerce (електронна търговия)

Самите източници предлагат и различни видове данни като: видео файлове, аудио файлове, текстови данни (съобщения и коментари), логове на данни (Sadiku, Adekunte, & Sadiku, 2024).

В сферата на телекомуникациите подходящи източници са електронните магазини на такива компании, защото чрез тях се запазва история на покупките и търсени компоненти. Друг подходящ източник е сайтът на компанията, където има данни за първата достъпена страница и всяка следваща до напускането на потребителя от сайта. Данни, свързани с посещението на клиентите във физическите магазини или дилърски магазини на компанията.

**Структура на данните**

След като вече бяха представени източниците на данни, може да бъде обърнато внимание и на структурата (Таблица 5) на самите данни: неструктурираните данни трябва да бъдат конвертирани в друг формат, позволяващ тяхното използване. Необходимо е да бъде намерен подход за тяхното конвертиране до структурирана или полу-структурирана форма, за да могат да бъдат анализирани и от тях да се извлекат знания, които да допринесат за решаването на актуални проблеми и взимането на бизнес решения.

*Таблица 5 Таблица съдържаща типовете структури на данните*

| **Структурирани** | **Полу-структурирани** | **Неструктурирани** |
|---|---|---|
| Данните са подходящи да постъпят директно в склада от данни на компанията. | Необходима е обработка, след която данните да могат да бъдат заредени в базата данни на компанията. | Не могат да бъдат заредени в склада от данни директно. |

Структурираните данни се отнасят до данни с определен формат и дължина, лесни за съхранение и анализ, с висока степен на нормализация. Това означава, че данните са организирани в разпознаваема структура с цел да се даде отговор на заявка или даден бизнес проблем. Типичен пример за структурирани данни е релационна база данни, която съдържа организирани числа, дати, групи думи и низове текст, а за език за формулиране на заявките - SQL (Rusu, O. & Halcu, Ionela & Grigoriu, O. & Neculoiu, Giorgian & Sandulescu, Virginia & Marinescu, M. & Marinescu, Viorel, 2013). Структурираните данни се характеризират с това, че могат да бъдат лесно организирани. Експертите в областта оценяват структурираните данни като само 20% от наличните данни.

## Откриване на скрити знания в данните - Data Mining

Извличането на знания от данни (Data Mining) става все по-популярно в съвременния свят. Заобиколени сме от огромни количества от данни, но без те да се превърнат в носеща смисъл информация, няма как да бъдат полезни. Откриването на смисъл в данните е възможно след като бъдат обработени и анализирани с помощта на различни data mining инструменти.

Според К. Стефанова и Д. Кабакчиева (Stefanova & Kabakchieva, 2015), с помощта на Data Mining в бизнеса успешно се решават проблеми, свързани с маркетинга, управлението на взаимоотношенията с клиенти, контрол на качеството при производство, управление на складовите наличности, финансово прогнозиране и т.н. В телекомуникациите ефективно се използват за запазване на клиенти, за оптимизация на телекомуникационните мрежи и откриване на проблеми в тяхното функциониране. В медицината методите за извличане на знания се прилагат ефективно за диагностика и намиране на подходящо лечение на базата на събрани исторически данни за пациентите. Data Mining методите намират приложение и в области като астрономия (например за откриване на нови космически обекти), в спорта (например за анализ поведението на играчите, дефиниране на игрова стратегия).

Главната цел при използването на Data Mining методите е извличането на смислена и полезна информация точно от тези големи обеми данни, като бъдат открити смислени взаимовръзки, зависимости и повтарящите се модели. Обработваните данни могат да бъдат в различни формати – цифров, текстови, изображения, времеви серии и други. В зависимост

от този формат, е необходимо да се изберат подходящи методи и алгоритми, които да бъдат използвани за обработка на данните.

**Типове Data Mining задачи**

Задачите за извличане на знания от данни (Data Mining) могат да бъдат разделени в два типа:

- Задачи за описание, където параметрите на данните се представят в обобщен вид. Към този тип задачи спада клъстерния анализ (cluster analysis);
- Задачи за предсказване, при които се разглеждат вече съществуващи данни, оценяват се и се обявяват прогнози въз основа на направените заключения.

Различават се два основни типа задачи за извличане на знания от данни – с обучение (directed) или без обучение (undirected). Извличане на скрити знания от данни, посредством обучение води категоризиране на определена целева променлива. Извличането на знания без обучение води до откриване на закономерности или общи черти сред групи от записи, без да се използват предварително зададена целева променлива или съвкупност от предварително определени класове (Stefanova & Kabakchieva, 2015).

**Задача за класификация**

За решаване на бизнес проблема с напускащите клиенти от сектора на телекомуникациите най-често се използва Data Mining задачата за класификация. Класификацията на данните представлява категоризирането на данни въз основа на разработени модели. При този тип задачи се предсказват стойности на категорийна променлива. Съществуват различни методи и алгоритми за класифициране при извличането на знания от данни, като например алгоритъм „Дърво на решенията“, „Невронни мрежи“, „Логистична регресия“ и т.н.

Идеята на моделирането е да се открият съществуващите взаимовръзки в изследваната съвкупност от данни. При описателното моделиране данните се обобщават с цел да бъдат разбрани по-добре данните и процесите, които биват описани. При предсказващото моделиране целта е да се предскажат правилно неизвестните стойности на избраната категорийна променлива, на базата на известните стойности на останалите променливи, които се съдържат в набора от данни.

Класифицирането на данни е процес в две стъпки:

- Първа стъпка - обучаването на алгоритъма, когато моделът бива създаван. Изгражда се класификатор, който работи с предварително известни класове (възможните стойности на избраната целева категорийна променлива) и определена точност на предсказване.
- Втора стъпка - използване на вече готовия модел върху нови данни с цел предсказване на класовете на принадлежност на обектите. Важно е да се отбележи, че ако точността на класификатора се определя от данните за обучение, той ще има

много висока успеваемост (Han, Kamber, & Pei, 2012), но няма да предсказва правилно по отношение на нови данни.

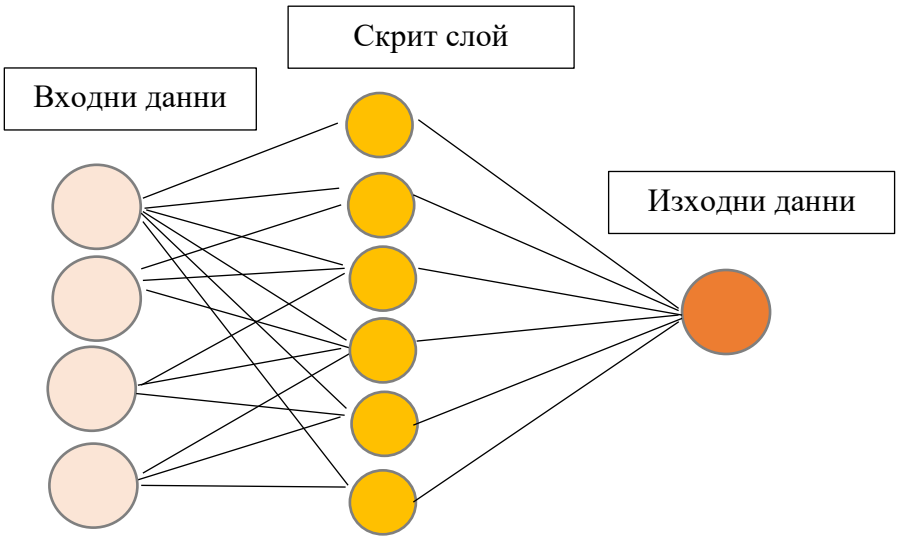
## 1.2. Обзор на методите за откриване на знания в данни (Data Mining) чрез обучение на класификатори

Както бе споменато по-горе, съществуват различни Data Mining методи за решаване на задачата за класификация, като най-често използваните са „Невронни мрежи“, „Дърво на решенията“, „Логистична регресия“ и др. Резултатите от извършения анализ на съществуващи изследвания на бизнес проблема с напускането на клиентите в телекомуникационния сектор, представени в Табл. 1, Табл. 2, Табл. 3 и Табл. 4 показват, че най-често използваните методи са „Невронни мрежи“, „Дърво на решенията“ и „Логистична регресия“, затова именно тези методи са представени и се използват в настоящето изследване.

**Невронни Мрежи**

Тъй като количеството от данни нараства всекидневно с огромни темпове, необходимостта от автоматизирана обработка става все по-голяма. Извличането на знания от данни позволява да бъде отговорено на въпроси, които все още не са зададени.

Невронните мрежи (Фиг 1) са един от най-често използваните методи за извличане на знания от данни. Вдъхновени са от невронните връзки на човешкия мозък. Този метод е приложим както за задачи от тип регресия, така и за класификационни задачи. Невронните мрежи са извести като адаптивен анализ, който може да промени структурата си по време на обучителния процес. След подаване на входните данни в скрития слой се създават сложни взаимовръзки между самите данни, с цел да се открият скрити модели в анализирания набор от данни. По своята същност, методът „Невронни мрежи“ се определя като нелинеен, вероятностен. Изборът как да бъде класифицирана дадена променлива, се случва посредством множество комбинации, базирани на теглата, необходими за активирането на връзките между невроните.




*Фигура 1 Визуална презентация на Невронна мрежа*

Чрез невронните мрежи данните, интерпретирани от различни източници, могат да бъдат обработени бързо и ефективно. За целта трябва да се направи модел на невронната мрежа, който се състои от три части: архитектура, алгоритъм за обучение и функции за активирането на правилата в него.

Изкуствените невронни мрежи (Artificial Neural Networks), се състоят от възли (nodes), разделени в различни слоеве. Входните данни се намират във входен слой, следват един или повече скрити слоеве и накрая е изходния слой, където се намират изходните данни. Всеки възел се свързва с другите възли като се формира тегло и праг, след достигането на който този възел бива активиран и изпраща данни до следващия слой на мрежата. В противен случай данните не се предават към следващия слой на мрежата.

За разлика от следващия метод, който ще бъде разгледан, при невронните мрежи обясняването на връзките между различните възли става далеч по-сложно, дори в някои случаи невъзможно, защото те са изчислителни елементи, които нямат определен бизнес смисъл.

**Дърво на решенията**

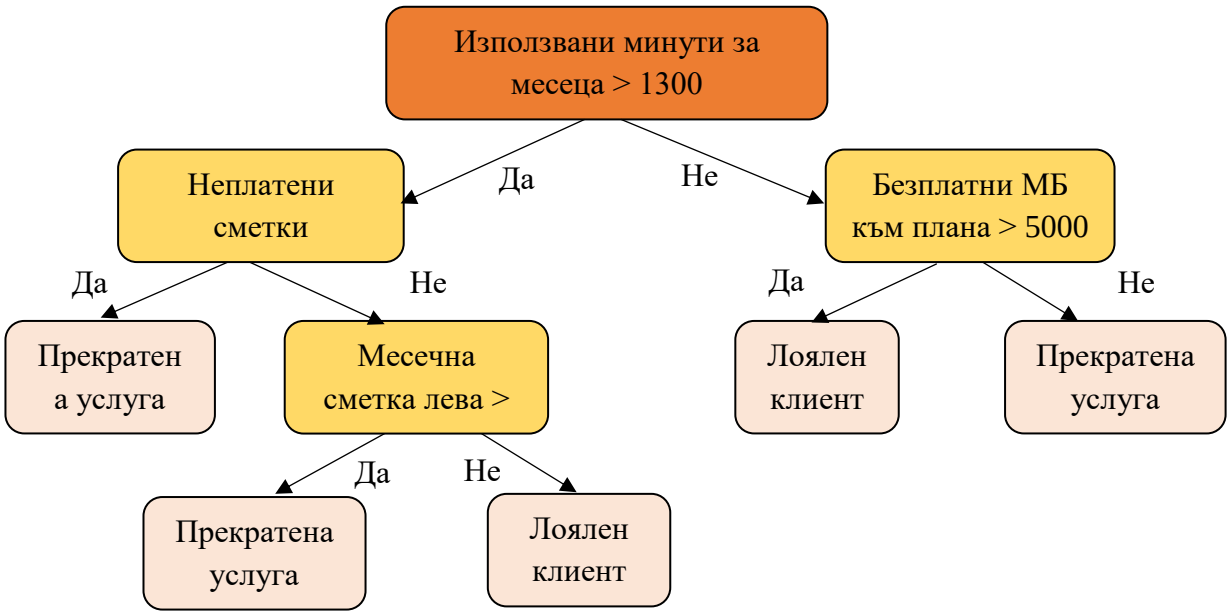
Друг често използван метод е „Дърво на решенията“, който също може да бъде използван за класификационни и регресионни задачи. Методът е еднакво подходящ както за числови, така и за категорийни данни. При прилагането на този метод се изгражда модел с дървовидна структура, включваща корен, клони и листа. Примерно дърво на решенията е показано на Фиг.2.

Коренът (root node) е основата на дървото, намираща се в най-горния му край, от тук започват разклоненията на „дървото на решенията“, които се определят от избраните променливи, участващи в конкретния модел.

Разклоненията се определят чрез резултатите от тестови процедури във вътрешните възли на дървото (decision nodes). Едно дърво може да позволява две или повече разклонения в зависимост от избрания алгоритъм за реализация.

Листата (leaf nodes) на дървото представят крайни възли, при които не може да има повече разклонения. Всяко листо съответства на някоя от стойностите на предсказваната променлива, примерно: отпадащи / задържащи се клиенти или решение: да/не.

Подрязване (Pruning) на дървото на решенията е термин, описващ намаляване на разклоненията с цел да се избегне прекалено нагаждане прилепяне (overfitting) спрямо данните за обучение.




*Фигура 2 Пример за Дърво на решенията*

По своята същност „Дърво на решенията“ е нелинеен, детерминистичен метод, използващ разделящи тестове от рода: вярно/невярно или тестове: Тест А/Тест Б.

При метода Дърво на решенията се използват различни критерии при разделящия тест - Entropy, Information gain, Gain ratio, Gini index, Accuracy.

**Логистична регресия**

Логистичната регресия е третия метод, който се използва в тази дисертация. Това е статистически метод, който използва предварително изследван набор от данни за обучение, с цел да предскаже стойностите на нов набор от данни. Логистичната регресия предлага много полезен анализ, когато задачите са свързани с категорийни данни. Това е един от най-използваните алгоритми за извличане на знания от данни, когато става въпрос за

класификация на дихотомни променливи (променливи с две стойности) - да или не, 1 или 0 (Witten, I. H. & Frank, E., 2005).

Логистичната регресия се използва за анализ на връзките между дихотомни зависими променливи и дихотомни независими променливи. Чрез нея могат да се комбинират независимите променливи, с цел да се провери вероятността да се случи определено събитие, т.е. целевата променлива да бъде член на една от групите, определени от дихотомната зависима променлива. Стойността от логистичната регресия е вероятност, която е в границите между 0 и 1. Ако вероятността целевата променлива да попадне в група, дефинирана от модела, е над определен праг, то целевата променлива става член на моделираната група. Когато вероятността е под зададения праг, то целевата променлива бива член на другата група. За всеки даден случай логистичната регресия изчислява вероятността данните от определен набор от атрибути, свързани с независимата променлива, да станат част от моделираната категория.

Необходимият размер на извадката за проучвания, свързани с наблюдение, които включват логистична регресия, е 500 записа според Буянг (Bujang MA, Sa'at N, Tg Abu Bakar Sidik TMI, Lim CJ, 2018), което условие е изпълнено при използваната съвкупност от данни.

## 1.3. Оценка и сравнение на получените Data Mining модели за класификация (обучените класификатори)

След получаването на Data Mining модели за класификация чрез избраните методи, тези модели трябва да бъдат оценени и сравнени.

Най-често използваните метрики за оценка на генерираните модели за класификация са Матрица на класификация (Confusion Matrix), Точност на класификацията (Classification Accuracy), ROC крива и др.

**Матрица на класификация (Confusion Matrix)**

Матрица на класификация представлява обобщение на точността на предсказаните резултати на определена задача за класификация.

Когато бива предсказвана бинарна променлива, класификационният модел поставя всеки възможен обект в един от двата възможни класа (True/False, Positive/Negative, 1/0). По този начин се формират 4 възможности за класификация на всеки обект. Тези възможности могат да бъдат представени в следната матрица (Таблица 7):

*Таблица 6 Матрица на класификацията*

| | **Real True** | **Real False** | |
|---|---|---|---|
| *Predicted True* | **TP (True Positive)** | **FP (False Positive)**<br>Type I error | **Precision**<br>TP/(TP+FP) |

| | | | |
|---|---|---|---|
| *Predicted False* | **FN (False Negative)**<br>Type II error | **TN (True Negative)** | **Negative Predictive Value**<br>TN/(TN+FN) |
| | **Sensitivity (Recall)**<br>TP/(TP+FN) | **Specificity**<br>TN/(TN+FP) | **Accuracy**<br>(TP+TN)/<br>(TP+TN+FP+FN) |

Метриките за оценка на точността на класификаторите се базират на стойностите в матрицата на класификация и включват:

Sensitivity – съотношение на правилно предсказаните позитивни резултати;

Precision – съотношението на правилно предсказаните позитивни резултати върху всички (позитивни и негативни) правилно предсказани резултати;

Accuracy (точност) – се изчислява от всички правилно предсказани резултати върху всички възможни резултати.

За финал може да бъде използвана оценка F1 score, която е средната претеглена стойност на sensitivity (recall) и precision. Тази оценка е добър избор, когато се търси баланс между изброените две.

$$F1\ Score = 2\frac{\text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Според Nikolaj Buhl в статията: „F1 Score in Machine Learning“ (Buhl, 2023), стойностите на F1 score показват следните нива на баланс:

- Под 0.5 – недобро ниво на баланс
- Между 0.51 – 0.8 – приемливо ниво на баланс
- Между 0.81 – 0.9 – добро ниво на баланс
- Над 0.91 – много добро ниво на баланс

**Точност на класификацията (Classification Accuracy)**

Най-разпространената метрика за изчисляване на точността на моделите за предсказване на база класификация се нарича Classification Accuracy (CA). Чрез CA може да бъде установено каква е производителността на модела като се разделят броя вярно

предсказани резултати върху всички предсказани резултати. Когато предсказаният клас от тестовата извадка съвпада с неговата реална стойност, се отчита правилно предсказване, а когато не съвпада, се отчита грешка.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Грешката при класификация представлява съотношение на броя обекти, които не са попаднали в правилния клас, разделен на общия брой обекти в тестовата извадка.

$$Error\ Rate = 1 - Accuracy.$$

Класификационната точност и грешка на моделите често се представят в проценти.

**ROC крива**

ROC кривата (**Receiver Operating Characteristic curve**) е често използвана мярка за визуализиране на производителността на задачи от тип класификация. Обикновено има два възможни изхода при така наречените бинарни класификатори – 0 и 1.

За оценка на точността на разработените класификатори, посредством различни Data Mining методи, се използва метриката AUC (Area Under the ROC Curve). Стойностите на метриката варират между 0 и 1. Колкото по-близо е стойността до 1, толкова по-точен е оценяваният класификационен модел. В таблица 8 са представени възможните стойности на AUC и тяхното тълкуване (Terra, 2025).

*Таблица 7 Тълкуване на стойностите на AUC*

| **Стойност на AUC** | **Значение** |
|---|---|
| *< 0.5* | По-зле от случайна класификация |
| *0.5- 0.6* | Близка до случайна класификация |
| *0.61 – 0.7* | Недобро представяне на модела |
| *0.71 – 0.8* | Адекватно представяне на модела |
| *0.81 – 0.9* | Добро представяне на модела |
| *0.91 – 1* | Отлично представяне на модела |

**Cohen’s Kappa coefficient**

Cohen’s Kappa coefficient (Kappa statistics) е метод за оценка на Data Mining модели, който се използва често от изследователите, търсещи скрити знания в данни. Мярката дава

възможност да се прецени каква е възможността предсказвания клас да съвпадне на случаен принцип с действителния клас. Коефициентът може да заема стойности в интервала -1 до 1:

- Коефициент = 1: означава пълно съвпадение на предсказвания клас спрямо реалния такъв;
- Коефициент = 0: означава, че предсказвания клас не е по-точен от клас избран на абсолютно случаен принцип;
- Коефициент = -1: означава, че предсказвания клас изобщо не съвпада с действителния такъв.

Тълкуване на Cohen's Kappa coefficient (Henry, Herwindiati, Mulyono, & Hendryli, 2016) (Таблица 9):

*Таблица 8 Тълкуване на Cohen's Kappa coefficient*

| **Стойност на Cohen's Kappa coefficient** | **Ниво на съгласуваност** |
|---|---|
| *0.01 – 0.2* | Ниско ниво |
| *0.21 – 0.4* | Достатъчно ниво |
| *0.41 – 0.6* | Умерено ниво |
| *0.61 – 0.8* | Високо ниво |
| *0.81 – 1* | Почти пълно съгласие |

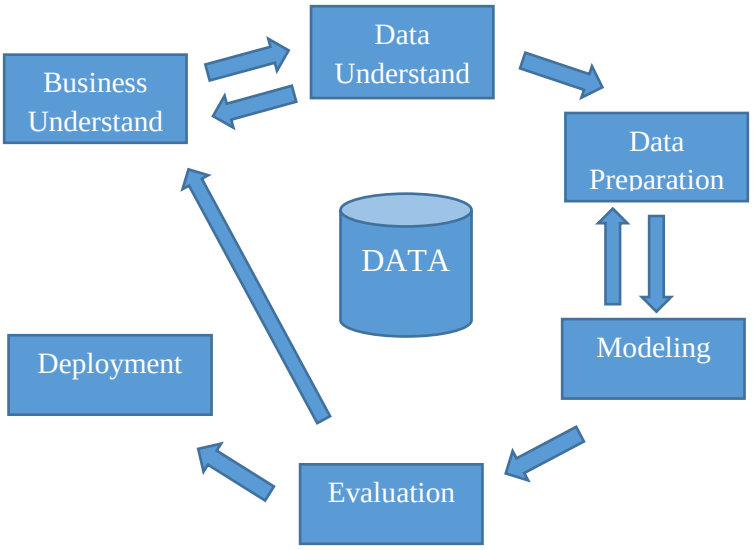
## 1.4. Съществуващи подходи за реализация на Data mining проекти (методология)

След избор на мерки за оценка на генерираните модели за класификация следва да се определи и подход за реализация на проекта.

Разгледани са най-често използваните методологии за реализация на Data Mining проекти: **CRISP-DM, KDD, SEMMA, Методология на 5-те А-та**. Избраният подход за реализация в този дисертационен труд е CRISP-DM, който е по-модерен от KDD и по-често използван от него в днешно време. Спрямо SEMMA подхода, CRIPS-DM е по-задълбочен, защото изисква познания на бизнес проблема, който бива решаван. Подходът на петте А отново засяга в малка степен бизнес частта и не е толкова гъвкав от гледна точка на промени в процеса на работа.

**CRISP - DM**

През 1999-2000г. методът CRISP-DM – Cross-Industry Standard Process for Data Mining е създаден от разработчици и потребители на софтуерни средства за Data Mining. Съгласно този модел, един цикъл на изследвания в областта на Data Mining може да се представи като процес, състоящ се от шест основни етапа (Фиг. 3). Последователността на етапите не е строго определена - често се налага връщане на предишен етап с цел подобряване на резултатите от следващ етап.




*Фигура 3 Визуална репрезентация на методът CRISP-DM*

*Разбиране на проблемната област (Business Understanding)*

Разбирането на проблемната област (Business Understanding) е началния етап, който се фокусира върху разбиране на целите на изследванията и формулиране на изисквания от гледна точка на бизнес потребителите.

*Разбиране на данните (Data Understanding)*

Този етап започва със събиране на необходимите данни и продължава с дейности, целящи задълбочаване на знанията на изследователя за естеството на данните.

*Подготовка на данните (Data Preparation)*

Подготовката на данните включва всички дейности по създаване от първоначални „сурови“ данни на „крайното множество“ от данни (т.е. данни, които ще бъдат използвани от моделиращите средства

*Моделиране (Modeling)*

Етапът моделиране (Modeling) включва избор и прилагане на различни методи за моделиране, целящи извличане на знания от данните.

*Оценка на моделите (Evaluation)*

Оценка на моделите (Evaluation) се прави с цел по- задълбочено разбиране на създадените модели от гледна точка не само на изследователя, но и на бизнес потребителите.

*Прилагане на моделите (Deployment)*

Готовите модели могат да се използват по два основни начина. Анализаторът може да препоръча предприемането на конкретни действия на базата на заключения от изградения модел и получените резултати, или моделът може да се приложи към нови данни.

## 1.5. Съществуващи софтуерни решения за реализация на Data Mining проекти

Следващата стъпка за осъществяването на Data Mining проекта е изборът на подходящ софтуерен инструмент за неговата реализация. Анализирани са следните четири инструмента: **WEKA**, **IBM SPSS Modeler**, **RapidMiner Studio** and **KNIME**.

**Сравнение на изброените инструменти**

След подробен анализ на изброените инструменти, е съставена сравнителна таблица с помощта, на която да бъде избран един от тях, за решаване на научния проблем в дисертационния труд.

Избрани критерии за сравнение на софтуерните инструменти за Data Mining:

- Лекота на използване;
- Гъвкавост на приложението;
- Потребителска общност;
- Извличане на знания от данни;
- Мащабируемост на приложението;
- Софтуер с отворен код.

От изброените Data Mining софтуери, за целите на изследването в дисертационния труд е избран RapidMiner Studio. Софтуерът разполага с богат набор от модули, модели и алгоритми, които да спомогнат за конструирането и изпълнението на задачата за откриване на напускащи клиенти в сферата на телекомуникациите.

Интерфейсът на инструмента е разбираем и улеснява потребителят му по време на работа. Също така, крайният резултат от модела може да бъде тълкуван с помощта на визуалните средства, предоставени от самият софтуер, без необходимост от допълнителни инструменти за съставяне на графики.

## 1.6. Изводи по Първа Глава

В първа глава на дисертацията е постигнато следното:

- ✓ Идентифициран е актуален бизнес проблем, свързан с напускането на клиенти в сферата на телекомуникациите, и е направено изследване относно най-често използваните аналитични методи за неговото решаване;
- ✓ Разгледани са основните източници на данни в сферата на телекомуникациите и за целите на дисертационния труд са избрани структурираните данни, които включват демографски характеристики на клиентите, данни генерирани от самите клиенти при използване на услугите на мобилния оператор и съхранявани в клиентска база данни на телекомуникационните компании, и данни свързани с метода на плащане, месечните и общите им плащания;
- ✓ Направен е обзор на актуалното състояние на научната област Data Mining (Извличане на скрити знания от данни), като са разгледани основните типове задачи и е поставен фокус върху задачата за класификация, тъй като именно тази Data Mining задача е подходяща за решаване на поставения бизнес проблем – навременно откриване на отпадащи клиенти;
- ✓ Направено е проучване относно най-често използваните Data Mining методи за решаване на бизнес проблема за откриване на отпадащи клиенти в сферата на телекомуникациите. За целите на дисертацията са избрани трите най-често използвани методи за класификация - „Невронни мрежи“, „Дърво на решенията“ и Логистична регресия;
- ✓ Предложени са различни метрики са оценка и сравнение на класификаторите, като за целите на дисертационния труд се използват Матрица на класификация (Confusion Matrix) и базираните на нея метрики за оценка, Точност на класификацията (Classification Accuracy), Площ под ROC кривата и Cohen’s Kappa coefficient;
- ✓ Представени са различни подходи за осъществяване на Data Mining проекти и за целите на дисертационния труд е избран подхода CRISP-DM;
- ✓ Разгледани са съществуващи софтуерни решения, подходящи за разработване на Data Mining модели за откриване на напускащи клиенти като, след сравнение, за целите на изследването в дисертационния труд е избран RapidMiner Studio;
- ✓ Формулирани са обект, предмет, хипотеза, цел и задачи на дисертационния труд.

# 2. Втора Глава: Методика за провеждане на изследването. Подготовка на данните.

Втора глава от настоящия дисертационен труд е съставена от три части. В първата част е представена методиката за провеждане на научното изследване. Във втората част на тази глава от дисертационния труд са включени описание и анализ на избраната съвкупността от данни, както и извършените дейности по предварителната подготовка на тези данни. В последната трета част на главата са представени изводите, свързани с методиката за провеждане на научното изследване и подготовката на данните, използвани за реализация на аналитичната задача.

## 2.1. Методика за провеждане на изследването

### Избор на подход за реализация на задачата за извличане на знания от данни (Data Mining) чрез обучение на класификатори

В този дисертационен труд се решава задачата за извличане на скрити знания от данни (Data Mining) посредством обучение на класификатори, с данни от сферата на телекомуникациите. За реализацията на Data Mining задачата за класификация е избран подходът CRISP-DM (Cross-Industry Standard Process for Data Mining), подробно описан в т.1.5.1.,защото той е най-често използвания подход за провеждане на Data Mining изследвания от специалисти в тази научна област.

Извличането на скрити знания от данни за потребители на компания от сферата на телекомуникациите е осъществено посредством последователно преминаване през първите пет етапа от подхода CRISP-DM.

В рамките на дисертационния труд, използвайки подходът CRISP-DM, е извършено следното на всеки от етапите:

- *Етап 1 – Разбиране на проблемната област*;
- *Етап 2 – Разбиране на данните*;
- *Етап 3 – Подготовка на данните*;
- *Етап 4 – Моделиране*;
- *Етап 5 – Оценка на моделите*;
- *Етап 6 – Прилагане на моделите*.

### Избор на софтуерни инструменти за провеждане на научното изследване

След изборът на подход за реализация на поставените задачи, е важно да бъдат избрани подходящи софтуерни инструменти, чрез които задачите да бъдат решени. В този дисертационен труд са необходими софтуерни инструменти първо за преглед и подготовка

на данните, след това за описание и изследване на данните, и финално за моделиране и оценка на разработените модели.

За целите на научното изследване, реализирано в този дисертационен труд, ще се използват:

- **Microsoft Excel** (Microsoft Office 2016) - за предварителен анализ и преглед на сетът от данни, извлечен от kaggle. С цел почистване на данните и тяхната подготовка във формат, подходящ за правилното им зареждa в Data Mining софтуер, за изчистване и предварителна подготовка на данните, за подготовка на данните във формат, подходящ за използване в Data Mining софтуер.
- **Power BI** (February 2025), **Microsoft Excel** (Microsoft Office 2016) и **RapidMiner Studio Free** (v9.9.002) за опознаване, изследване и описание на данните.
- **RapidMiner Studio Free** (v9.9.002) - за моделиране и оценка на получените резултати.

**Описание на методиката за провеждане на научното изследване**

Научното изследване, което е представено в този дисертационен труд, е осъществено посредством изпълнението на следните стъпки:

- **Проучване на литературни източници**;
- **Подбор на данните от избраната приложна област, изучаване на същността, произхода, начина на събирането, организирането и съхранението им;**
- **Изучаване и предварителна подготовка на избраната съвкупност от данни**;
    - **Визуален анализ на данните чрез софтуерните инструменти Excel, Power BI и RapidMiner Studio**.
- **Разработване на Data Mining модели за класификация**;
- **Подобряване на Data Mining моделите за класификация**;
- **Оценка и сравнение на разработените Data Mining модели за класификация:** Избраните критерии за оценка и сравнение са както следва:
    - Classification Accuracy – Точност на класификацията – процентно отношение на правилно предсказаните обекти върху всички обекти;
    - Error Rate – Грешка на класификацията - процентно отношение на грешно предсказаните обекти върху всички обекти;
    - Kappa Statistic (Cohen’s Kappa) – Критерии, чиято стойност показва степен на съгласуваност между предсказания и действителния клас;
    - Confusion Matrix – Матрица на класификацията – Съдържа информация за резултатите от направената класификация;

  - Параметрите F-1 Score, Precision и Recall, изчислени на базата на стойностите в Матрица на класификацията;
- **Изводи и препоръки към потенциални крайни потребители, интересуващи се от резултатите на демонстрираните Data Mining анализи**.

## 2.2 Разбиране и подготовка на данните

Научното изследване, представено в настоящата дисертация, е осъщественото върху данни за клиенти на компания от сферата на телекомуникациите. Данните съдържат демографски характеристиките на потребителите, информация за използваните от тях услуги, начини на плащане и платени сметки.

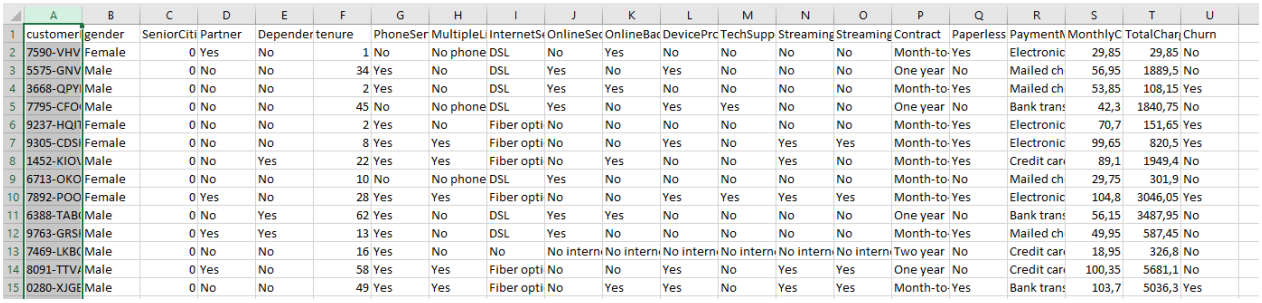
Обобщено описание на избраната съвкупност от данни, използвана в дисертацията:

- Източник на данните – данните могат да бъдат открити в софтуерния инструмент Orange или в базата от данни на Kaggle (един от най-големите източници на свободно достъпни сетове от данни);
- Формат на данните – форматът на данните е CSV (Comma Separated Values), един файл, една електронна таблица;
- Обем на данните – данните включват 7043 записа, общо 21 променливи. Двайсет от променливите са описателни и една променлива за предсказване.

**Подготовка на данните**

Подготовката на данните е третия етап от избрания CRISP-DM подход за провеждане на научното изследване, като това е много важен и продължителен процес. Качеството на разработените модели до голяма степен зависи от подходящата предварителна подготовка на данните.

Правилното визуализиране на данни е извършено след като успешно е разпознат десетичния делител на числовите данни. Пропускането на тази стъпка, при визуализирането на избрания сет от данни, води до грешна интерпретация на числови данни, представяйки ги като дати. След прилагане на подходящия делител съвкупността от данни се визуализира готова за последващи анализи (фигура 4).



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | customer | gender | SeniorCiti | Partner | Depender | tenure | PhoneSer | MultipleL | InternetS | OnlineSec | OnlineBac | DevicePro | TechSupp | Streaming | Streaming | Contract | Paperless | PaymentM | MonthlyC | TotalChar | Churn |
| 2 | 7590-VHV | Female | 0 | Yes | No | 1 | No | No phone | DSL | No | Yes | No | No | No | No | Month-to | Yes | Electronic | 29,85 | 29,85 | No |
| 3 | 5575-GNV | Male | 0 | No | No | 34 | Yes | No | DSL | Yes | No | Yes | No | No | No | One year | No | Mailed ch | 56,95 | 1889,5 | No |
| 4 | 3668-QPY | Male | 0 | No | No | 2 | Yes | No | DSL | Yes | Yes | No | No | No | No | Month-to | Yes | Mailed ch | 53,85 | 108,15 | Yes |
| 5 | 7795-CFO | Male | 0 | No | No | 45 | No | No phone | DSL | Yes | No | Yes | Yes | No | No | One year | No | Bank trans | 42,3 | 1840,75 | No |
| 6 | 9237-HQI | Female | 0 | No | No | 2 | Yes | No | Fiber opti | No | No | No | No | No | No | Month-to | Yes | Electronic | 70,7 | 151,65 | Yes |
| 7 | 9305-CDS | Female | 0 | No | No | 8 | Yes | Yes | Fiber opti | No | No | Yes | No | Yes | Yes | Month-to | Yes | Electronic | 99,65 | 820,5 | Yes |
| 8 | 1452-KIO | Male | 0 | No | Yes | 22 | Yes | Yes | Fiber opti | No | Yes | No | No | Yes | No | Month-to | Yes | Credit car | 89,1 | 1949,4 | No |
| 9 | 6713-OKO | Female | 0 | No | No | 10 | No | No phone | DSL | Yes | No | No | No | No | No | Month-to | No | Mailed ch | 29,75 | 301,9 | No |
| 10 | 7892-POO | Female | 0 | Yes | No | 28 | Yes | Yes | Fiber opti | No | No | Yes | Yes | Yes | Yes | Month-to | Yes | Electronic | 104,8 | 3046,05 | Yes |
| 11 | 6388-TAB | Male | 0 | No | Yes | 62 | Yes | No | DSL | Yes | Yes | No | No | No | No | One year | No | Bank trans | 56,15 | 3487,95 | No |
| 12 | 9763-GRS | Male | 0 | Yes | Yes | 13 | Yes | No | DSL | Yes | No | No | No | No | No | Month-to | Yes | Mailed ch | 49,95 | 587,45 | No |
| 13 | 7469-LKB | Male | 0 | No | No | 16 | Yes | No | No | No intern | No intern | No intern | No intern | No intern | No intern | Two year | No | Credit car | 18,95 | 326,8 | No |
| 14 | 8091-TTV | Male | 0 | Yes | No | 58 | Yes | Yes | Fiber opti | No | No | Yes | No | Yes | Yes | One year | No | Credit car | 100,35 | 5681,1 | No |
| 15 | 0280-XJG | Male | 0 | No | No | 49 | Yes | Yes | Fiber opti | No | Yes | Yes | No | Yes | Yes | Month-to | Yes | Bank trans | 103,7 | 5036,3 | Yes |

*Фигура 4 Визуален изглед на данните след прилагане на правила за разделяне на колоните и обработка на числовите стойности*

В избраната съвкупност от данни има общо 21 променливи, една целева променлива – Churn, и 20 описателни променливи - четири от тях са числови, а другите шестнайсет са категориини. Единствено променливата SerniorCitizen, която е индикатор дали възрастта на потребителя (под и над 65 години), изглежда с описателен характер, но е представена чрез числови стойности.

Важна стъпка от подготовката на данните е проверката за корелация между променливите. В случай на откриване на висока корелация, е препоръчително една от двойката променливи да отпадне от разработката на модела за откриване на скрити знания в данни.

Използваните методи, за извличане на скрити знания от данни в дисертационния труд са „Дърво на решенията“, „Невронни мрежи“ и „Логистична регресия“. При използването на нелинейни методи, каквито са „Дърво на решенията“ и „Невронни мрежи“, високата корелация не влияе на резултатите негативно, както при линейните методи (Chowdhury, Lin, Liaw, & Kerby, 2021). При „Логистична регресия“, който по своята същност е линеен метод, корелация на променливите над 85% влияе негативно на точността на предсказване, но обикновено границата, над която една от корелираните променливите бива премахнат, при разработване на модели, е 80% (Vatcheva, Lee, McCormick, & Rahbar, 2016). Допълнителна проверка за корелация е извършена в трета глава на дисертационния труд т.3.3.

**Изследване и описание на данните**

Крайната съвкупност от данни (Kaggle, 2016), използвана в дисертацията, е описана в Табл. 12:

*Таблица 9 Наименование, описание, типове, възможни стойности и липсващи стойности на променливите, съставящи избраната съвкупност от данни.*

| Номер | Променлива | Описание | Тип на променливата | Възможни стойности | Липсващи стойности |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| 1 | customerID | Уникален идентификатор на потребителя | Категорийна | Всяка от стойностите е уникална | не |
| 2 | Gender | Пол на потребителя | Категорийна | Female, Male | не |
| 3 | SeniorCitizen | Индикатор дали възрастта на потребителя е над 65 | Числова | 0 = под 65 години, 1 = над 65 години | не |
| 4 | Partner | Показва дали клиентът има партньор или не | Категорийна | Yes = да, No = не | не |
| 5 | Dependents | Индикатор дали потребителя живее с зависими от него лица (деца, родители, възрастни хора и други) | Категорийна | Yes = да, No = не | не |
| 6 | Tenure | Показва общия брой месеци, в които потребителят е бил клиент на компанията. | Числова | От 0 до 72 месеца | не |
| 7 | PhoneService | Показва дали клиентът има услуга за стационарен телефон. | Категорийна | Yes = да, No = не | не |
| 8 | MultipleLines | Показва дали клиентът има повече от един мобилен/стационарен план. | Категорийна | Yes = да, No phone = не | не |
| 9 | InternetService | Показва дали клиентът има услуга интернет. | Категорийна | No = без интернет план, DSL, Fiber Optic, Cable = разновидности на интернет плановете | не |

| | | | | | |
|---|---|---|---|---|---|
| 10 | OnlineSecurity | Показва дали клиентът има допълнителна услуга за интернет защита. | Категорийна | Yes = да, No = не, No internet service = без интернет услуга | не |
| 11 | OnlineBackup | Индикатор дали клиентът има допълнителни бекъп услуги предоставяне от компанията | Категорийна | Yes = да, No = не, No internet service = без интернет услуга | не |
| 12 | DeviceProtection | Индикатор дали клиентът има допълнителна застраховка на интернет устройствата си, предоставени от компанията. | Категорийна | Yes = да, No = не, No internet service = без интернет услуга | не |
| 13 | TechSupport | Индикатор дали клиентът има абонамент за техническа помощ, предоставена от компанията. | Категорийна | Yes = да, No = не, No internet service = без интернет услуга | не |
| 14 | StreamingTV | Съдържа информация дали клиентът използва своята интернет услуга за да гледа телевизионни канали от трети страни (не се начислява допълнително). | Категорийна | Yes = да, No = не, No internet service = без интернет услуга | не |
| 15 | StreamingMovies | Съдържа информация дали клиентът използва своята | Категорийна | Yes = да, No = не, No internet service = без интернет услуга | не |

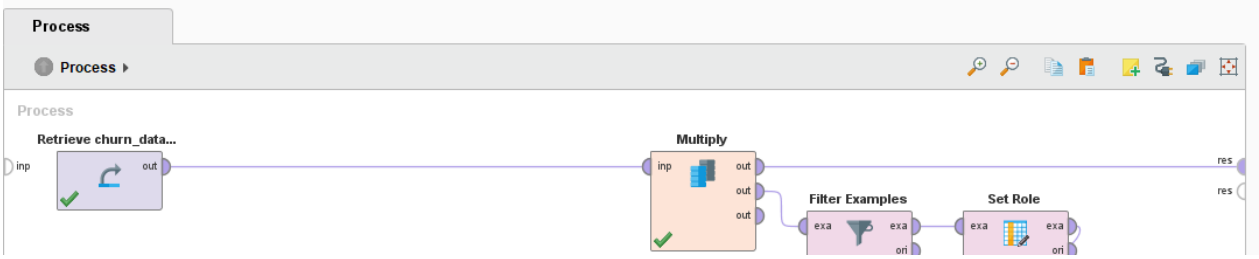
| | | | | | |
|---|---|---|---|---|---|
| | | интернет услуга за да гледа филми предоставени от трети страни (не се начислява допълнително). | | | |
| 16 | Contract | Показва каква е продължителността на договора на клиента | Категорийна | Month-to-month = месечно заплащане без годишен договор, One year = едногодишен договор, Two year = двугодишен договор | не |
| 17 | PaperlessBilling | Показва дали клиентът е избрал да не получава хартиена фактура. | Категорийна | Yes = да, No = не | не |
| 18 | PaymentMethod | Показва как клиентът плаща своите сметки | Категорийна | Electronic check = плащане онлайн,<br>Mailed check = плащане по поща,<br>Bank transfer (automatic) = автоматично банково плащане,<br>Credit card (automatic) = автоматично плащане с кредитна карта | не |
| 19 | MonthlyCharges | Съдържа месечните разходи за | Числова | От 18.25 до 118.75 долара, средна | не |

| | | | | | |
|---|---|---|---|---|---|
| | | всички услуги на клиента | | аритметична стойност: 64.76, медиана: 70,35, мода: 20.20 | |
| 20 | TotalCharges | Съдържа общите разходи за клиента през цялото време докато е клиент на компанията | Числова | От 0 до 8684.80 долара, средна аритметична стойност: 2283.30, медиана: 1397.45, мода: 20.05 | Да – за 11 потребителя |
| 21 | Churn | Дава информация дали клиентът е прекратил своите услуги или не | Категорийна | Yes = да, No = не | не |

**Премахване на записи с липсващи стойности**

След направеното изследване на данните е установено, че за някои променливи има липсващи стойности. Това може да бъде потвърдено и чрез проверка в Excel, използвайки следната функция „=COUNTBLANK(range)“ и по-конкретно „=COUNTBLANK(A2:A7044)“. Функцията е приложена върху първата променлива, намираща се в колона A, и след това върху всички останали променливи. Проверката потвърждава 11 липсващи стойности само при променливата TotalCharges.

Проверката може да бъде извършена и в RapidMiner Studio, като наборът от данни бъде свързан към резултатната визуализация на процеса (Фиг. 6).




*Фигура 5 Изглед на процеса, направен в RapidMiner Studio, с цел визуализация на основните характеристики на данни*

Общото между тези 11 липсващи стойности е, че клиентите са от 0 месеца в компанията. Това води до заключението, че те не са заплатили първите си фактури. Друга

обща черта е, че въпросните клиенти имат показател „Yes“ в колона Dependents. Трета обща черта на тези 11 клиента е, че никой от тях не е маркиран като възрастен гражданин (над 65 години) в колона SeniorCitizen. Според данните в променливата Churn, която се използва като целева променлива за предсказване в изследването, горните клиенти са отбелязани със стойност „No“. Тези потребители са част от списъка с клиенти на компанията, но изглежда, че те са се отказали от своите услуги веднага и не са маркирани като отпаднали. За целите на настоящия дисертационен труд, тези 11 клиента са премахнати чрез специален оператор при разработването на Data Mining модела в RapidMiner Studio. Установено е, че данните за потреблението на услугите за тези 11 записа са непълни, следователно биха допринесли негативно за точността на разработваните модели за класификация.

**Изследване на стойностите на променливите**

Честотното разпределение на предсказваната променливата е направено с помощта на софтуер Power BI.

В общата съвкупност от данни има 1869 (26,5%) потребители, които са прекратили услугите си (оцветени в червено) и 5174 (73,4%), които са продължили да бъдат клиенти на компанията.

Направен е анализ на демографските характеристики на клиентите в общата съвкупност от данни. Присъствието на клиенти от двата пола, жени и мъже, в данните е почти по равно. Мъжете са 3555 (50,48%), а жените 3488 (49,52%). Също така, почти половината от изследваните клиенти имат партньор – 3402 (48,30%), а малко над половината 3641 (51,70%) в момента нямат такъв. В сета от данни предимно има хора под 65 години 5901 (83%), а възрастни хора над 65 годишна възраст са 1142 (16%). Една трета от клиентите имат човек, който зависи от тях – 2110 (30%), а 4933 (70%) потребителя се грижат основно за собствените си нужди.

Анализ на услугите, предоставени от компанията. От извършения анализ става ясно, че повечето потребители имат включена услугата за телефонни разговори. Най-голям дял от клиентите имат активиран оптичен интернет – 3096 (44%). Другите допълнителни услуги са активирани от по-малко от 50% от клиентите, участващи в сета от данни.

Останалата част от категорийните променливи е свързана с дължината на подписания договор, типа и метода на плащане на потребителите:

- Contract;
- PaperlessBilling;
- PaymentMethod.

Повече от половината клиенти са подписали услуга, обвързана с месечен вместо годишен договор, и са избрали електронни фактури. Най-голям дял от плащанията биват

извършени по електронен път. Повечето клиенти (59,22%) предпочитат да не получават хартиена фактура за сметката си.

От числовите данни в сета от данни се наблюдават следните променливи:

- Tenure;
- MonthlyCharges;
- TotalCharges.

## 2.3. Изводи по Втора Глава

Във втората глава на дисертацията са постигнати следните резултати:

- ✓ Представена е методиката за провеждане на научното изследване, включваща:
  - Избрания подход CRISP-DM за решаването на Data Mining задачата за класификация, като са описани основните етапи, през които преминава реализираното изследване;
  - Избраните софтуерни инструменти, които се използват за реализирането на поставените задачи. Excel и PowerBI се използват за изследване и описание на данните. Предварителната подготовка на данните е извършена в Excel. За решаването на Data Mining задачата за класификация се използва софтуерния инструмент RapidMiner Studio – за създаване на класификаторите и за оценка на тяхната точност на предсказване;
  - Основните стъпки от методиката, които се осъществяват за реализацията на научното изследване.
- ✓ Описани са данните, избрани и използвани за целите на дисертационния труд, както и предприетите дейности по предварителната подготовка на данните.
- ✓ Финалната съвкупност от данни включва 7032 записа на потребители на телекомуникационна компания. Сетът от данни разполага с 20 описателни променливи и 1 една изходна променлива, която представлява целевата променлива за предсказване;
- ✓ Предсказваната променлива е „Churn“ със стойности „Yes“ – отпаднал клиент и „No“ – клиент, който все още не е прекратил договорът си.
- ✓ Проучване на литературни източници в сферата на извличането на скрити знания от данни;

# 3. Трета Глава: Резултати от изследването на моделите, получени чрез Data Mining задачата за класификация

В ***Трета глава*** са представени резултатите от изследването, осъществено чрез прилагане на избраните Data Mining методи и алгоритми за класификация върху предварително подготвената крайна съвкупност от данни за клиентите на телекомуникационна компания, с помощта на Data Mining софтуер RapidMiner Studio Оценени, подобрени и сравнени са генерираните Data Mining модели за класификация (обучените класификатори) чрез избраните метрики за оценка. Представени са резултатите за получените класификатори по метода „Дърво на решенията“, „Невронна мрежа“ и по метода „Логистична регресия“. Направено е допълнително сравнение на резултатите от подобрените модели и са представени изводите от трета глава.

Софтуерният инструмент, който се използва в дисертационния труд за реализация на Data Mining задачата за класификация, е RapidMiner Studio.

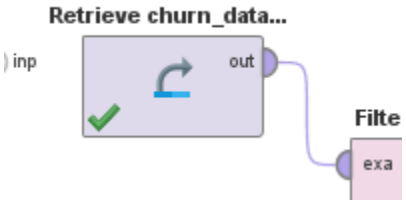
Преди да бъдат представени работните процеси за разработване на моделите за класификация, е важно да се отбележи, че част от началните стъпки участват и при трите използвани метода. Поради тази причина, повтарящите се стъпки са описани предварително и са валидни за работните процеси при трите метода, а индивидуалните стъпки за всеки метод са представени съответно в т. 3.1, т. 3.2 и т.3.3.

Основни повтарящи се стъпки:

- Стъпка I – Зареждане на предварително избраните данни, като се избира техният източник (локален или достъпен в интернет пространство);
- Стъпка II – Филтриране на липсващи или грешни стойности;
- Стъпка III – Избор на променлива за предсказване;
- Стъпка IV – Избор на описателни променливи, които да участват в модела.

**I. Зареждане на данните**

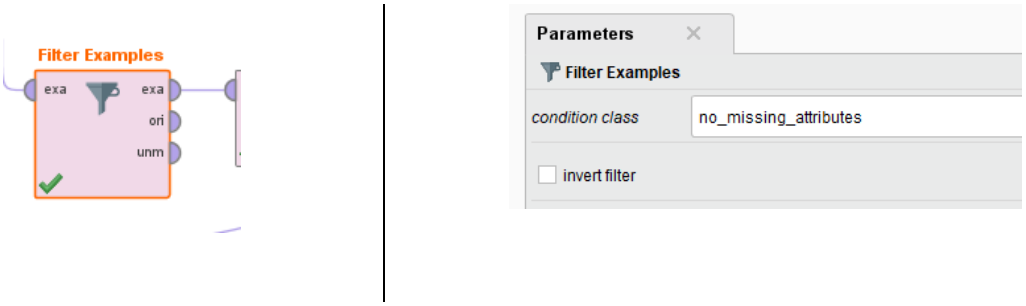
Първо трябва да бъдат заредени данните, които ще бъдат използвани.




*Фигура 6 Оператор за зареждане на данните в процес направен в RapidMiner Studio*

### II. Филтриране на непълните данни

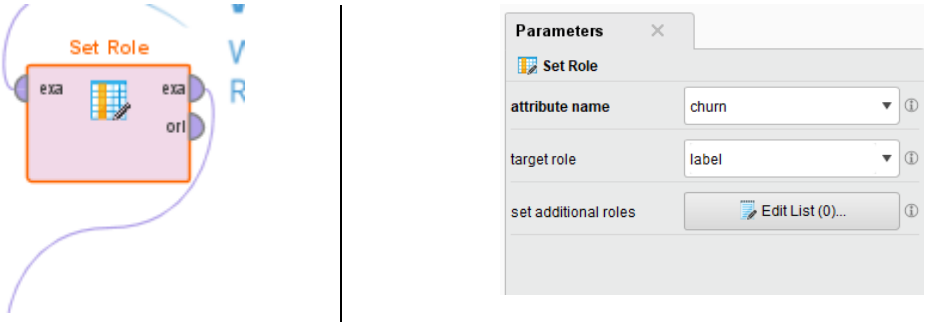
След това е необходимо да бъдат премахнати или допълнени данни, които биха повлияли на модела негативно. Това вече е направено на по-ранен етап от научното изследване, но тъй като използваният инструмент има такава функционалност и това е важна стъпка от методологията за създаване на подходящ Data Mining модел, на Фигура 7 е даден пример как е възможно да се направи това чрез оператора „Filter Examples“, като бъде избрано условието condition class = no missing attributes.




*Фигура 7 Оператор за филтриране на записите участващи в процеса*

### III. Избор на променлива за предсказване

Следва стъпка Set Role, тук е необходимо да бъде посочена изследваната за предсказване променлива и нейната роля - това е категорийната променлива „Churn“ (target role - label).




*Фигура 8 Оператор за определяне на целевата променлива за предсказване*

### IV. Избор на описателните променливи, които ще участват в модела

Много важно е да се провери има ли корелация между описателните променливите в използваната съвкупност от данни, спрямо променливата за предсказване. След като бъде установена такава, част от променливите се премахват, защото те се описват взаимно едни други, т.е. информацията вече е налична в някоя от другите характеристики. Най-добра променлива е тази с най-малък коефициент на корелация. Корелацията обаче не винаги е най-добрият метод за определяне на подходящите променливи. В дисертационния труд са

разгледани следните 3 варианта за оптимален подбор на променливите, които участват в модела:

- Weights by Correlation
- Weights by Information Gain
- Weights by Information Gain Ratio

***Weights by Correlation***: При този метод на претегляне на променливите, най-добрите резултати са свързани с по-малките коефициенти (1 - абсолютна корелация, 0 – не съществува никаква връзка между променливите.

Променливите с най-голяма корелация са: „Contact“, „Tenure“, „OnlineSecurity“ и „TechSupport“, но след като вече е преминала втора стъпка от подхода за реализиране на проекта (data understanding) е установено, че променливата „Tenure“ не може да бъда изключена като описателна/входна характеристика, защото е значима за откриването на напускащите клиенти.

***Weights by Information Gain***: Information Gain се използва за оценка на релевантността на променливите в определена съвкупност от данни като измерва количеството информация, която описателните променливите предоставят за предсказваната променлива. При този метод променливите с по-висок коефициент са за предпочитане.

Отново, сравнявайки коефициентите на атрибути със значимостта на данните се вижда, че и този подход е неподходящ. Променливата „CustomerID“ е определена като най-значима, а сама по себе си тази променлива не носи никаква добавена стойност за откриването на напускащите клиенти, тъй като всеки клиент има уникален номер и тази променлива не може да се използва за намиране на сходство.

***Weights by Information Gain Ratio***: Третият метод за подбор на променливите, които ще участват в модела, води до получаване на най-добри резултати.

Отново методът изчислява релевантността на променливите в съвкупността от данни като измерва количеството информация, която описателните променливите предоставят за предсказваната променлива, но същевременно коригира пристрастията към променливи с много възможни стойности. При Weights by Information Gain Ratio по-високите коефициенти са свързани с по-добри резултати. В този случай важни променливи като „TotalCharges“, „Contract“, „Tenure“, „TechSupport“, „InternetService“ са с по-висок коефициент.

Изборът на променливите, които да останат в модела, се случва в “Select by Weights” оператора. Зададена е граница 0.04 на коефициентите, като тя позволява да останат 10 променливи в модела. На пръв поглед повече на брой променливи могат да доведат до по-добри крайни резултати, но това е свързано с „прилепване“ към обучаващите данни (overfitting), което ще направи модела по-неточен по отношение на нови данни. Също така,

при решаване на реален сходен проблем, използването на множество различни описателни променливи, които да бъдат попълнени за всеки запис, става почти невъзможно.

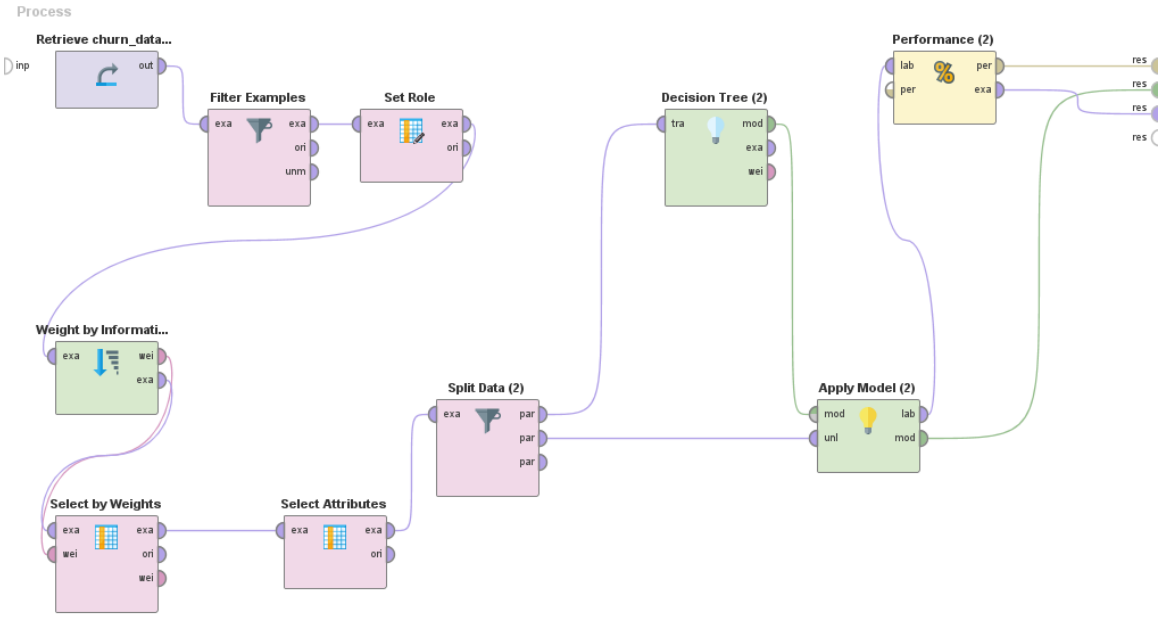
Финалният набор от данни, който се използва по-нататък за целите на изследването, включва 9 описателни променливи - 3 от тях са от числов тип, 6 са от категориен тип, и предсказваната променлива, която също е категорийна. Променливите, които ще участват в при разработването на Data Mining моделите са:

- Tenure
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- Contract
- MonthlyCharges
- TotalCharges
- Churn

## 3.1. Генериране, оценка и подобряване на Data Mining класификационен модел с метода „Дърво на решенията“

Първият изготвен модел за откриване на напускащи клиенти в сферата на телекомуникациите, представен в тази разработка, е чрез метода „Дърво на решенията“. При този метод се изпълняват поредица от класификационни правила, като по този начин всеки обект от избраната съвкупност от данни попада в определен клас.

Цялостен изглед на процеса е представен на фигура 9:




*Фигура 9 Цялостен изглед на процеса създаден със софтуер RapidMiner*

Основните стъпки в процеса за разработване на Data Mining модел, чрез метода „Дърво на решенията“ са:

- Стъпка I – Зареждане на предварително избраните данни, като се избира техният източник (локален или достъпен в интернет пространство);
- Стъпка II – Филтриране на липсващи или грешни стойности;
- Стъпка III – Избор на променлива за предсказване;
- Стъпка IV – Избор на описателни променливи, които да участват в модела;
- Стъпка V – Разделяне на данните на две групи, за обучение и за тестване;
- Стъпка VI – Избор на Data Mining метода, който ще се използва за решаването на Data Mining задачата;
- Стъпка VII – Прилагане на Data Mining метода върху данните за обучаване;
- Стъпка VIII – Оценка на получения класификационен модел (класификатор) върху тестовите данни
- Стъпка IX – Подобряване на процеса с цел получаването на класификатор с по-висока точност

Следва описание на стъпките V-IX, които са специфични за метода „Дърво на решенията“, тъй като първите четири стъпки, идентични за всички методи, са представени в началото на трета глава на дисертационния труд.

**V. Разделяне на данните (за обучение и тестване на получения модел)**

След това е необходимо наборът от данни, който се използва, да бъда разделен на две части - данни, които ще послужат за обучаването на модела и такива, на които ще бъде

приложен моделът, за да се оцени неговата точност. 70% от данни ще бъдат използвани за обучение, а на останалите 30% моделът ще бъде изпробван. Това съотношение се използва най-често при работа с метода „Дърво на решенията“ и при всички останали методи, използвани в дисертационния труд.

Операторът „Split Data“ може да се използва, за да се разделят данните на подмножества, като софтуерът предлага няколко вида на делене:

- ***Linear sampling***: Разделя сета от данни на части, без да променя реда на примерите, като създава подмножества с последователни редове;
- ***Shuffled sampling***: Изгражда случайни подмножества от сета данни. Примерите се избират на случаен принцип за създаване на подмножества;
- ***Stratifed sampling***: Изгражда случайни подмножества и гарантира, че разпределението на класовете в подмножествата е същото като в целия сет от данни.

**VI. Избор на Data Mining метод**

В тази стъпка се определя Data Mining методът, който ще бъде използван. При метод „Дърво на решенията“ има възможност за задаване на стойностите на различни параметри на алгоритъма, като:

- Критерий(criterion), който определя как да бъде извършено разделянето във вътрешните възли на дървото. В случая е избран - Gain_ratio;
- Максимална дълбочина (maximal depth): максималната дълбочина на дървото на решенията е зададена като ограничение в броя разклонения, които могат да се случат при изграждането му, в случая е избрана стойност 10;
- Доверителен интервал (confidence): 0.05 или 5%;
- Minimal leaf size: минимален брой обекти, попадащи в едно листо, в случая е избрана стойност 20
- Minimal size for split: предварително определен минимален брой обекти, попадащи в едно разклонение, за да съществува такова, в случая е избрана стойност 4;
- Number of Prepruning alternatives: спира растежа на дървото на решенията преди перфектно да класифицира тренировъчните данни. Предпазва от „нагаждане“ /„прилепване“ (overfitting), в случая е зададена стойност 3.

**VII. Прилагане на избрания алгоритъм върху обучаващите данни**

За целта към оператора „Decision Tree“ се подават обучаващите данни, които представляват 70% от общата съвкупност от данни.

**VIII. Оценка на получения модел (класификатор) върху тестовите данни**

Последната стъпка е свързана с проверката на успеваемостта на модела. Полученият модел (класификатор) се прилага върху тестовата извадка от данните (останалите 30% от данните), които не са участвали в генерирането на модела, за да се оцени неговата точност.

За целта към оператора „Apply Model се подават тестовите данни и полученият класификатор на изхода на „Decision Tree“ оператора.

Изборът на метрики за оценка на модела се осъществява в операторта „Performance“. За оценка на модела може да се използват повече метрики, но по подразбиране е избрана опцията Accuracy, която представлява процентно съотношение на правилно предсказаните променливи спрямо всички налични.

Получените резултати от оценката на модела са представени на Фиг. 10.

**accuracy: 78.15%**

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 1348 | 260 | 83.83% |
| pred. Yes | 201 | 301 | 59.96% |
| class recall | 87.02% | 53.65% | |

*Фигура 10 Матрица на класификацията, получена от прилагането на Метода „Дърво на решенията“*

Първоначалните резултати са получени чрез алгоритъм по метода „Дърво на решенията“ със следните настройки, представени в Таблица 14:

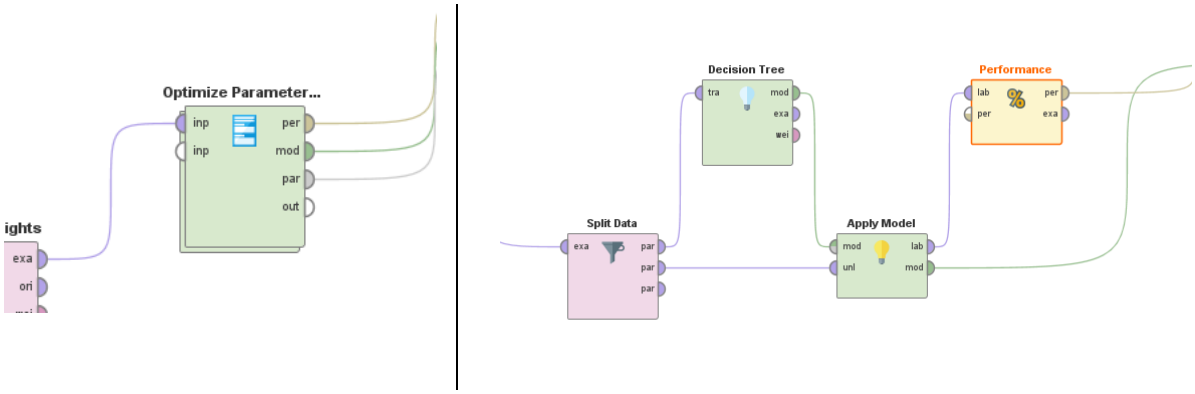
*Таблица 10 Първоначални настройки на модела, изпълнен с метод „Дърво на решенията“*

| **Наименование на параметрите в модела** | **Стойност** |
|---|---|
| Criterion | Gain ratio |
| Maximal Depth | 10 |
| Minimal leaf size | 20 |
| Confidence | 0.05 |

Както се вижда на Фиг. 10, цялостната точност на предсказване на получения класификатор (предсказване на двата класа) е 78.15%. Точността на предсказване на класа на отпадащите клиенти е 59.96%.

## IX. Подобряване на процеса

Оптимизацията на процеса се извършва с оператора „Optimize Parameters (Grid)“. В него се поставя частта от процеса след избора на описателните променливи (Фиг. 11).




*Фигура 11 Оператор за оптимизация на процеса и съдържащи се в него стандартни оператори*

В Табл. 15 са представени настройките за оптимизиране на процеса в оператора „Optimize Parameters (Grid)“.

*Таблица 11 Оптимизиране на настройки на модела, изпълнен с метод „Дърво на решенията“*

| **Наименование на параметрите в модела** | **Възможни стойности** |
|---|---|
| Criterion | Gain ratio, Information gain, Gini Index, Least Square |
| Maximal Depth | Между 1 и 100, със стъпка 10 |
| Minimal leaf size | Между 1 и 100, със стъпка 10 |
| Split data – Sampling type | Automatic, shuffled, stratified и linear |

Има възможност да бъдат избрани повече от един критерий, с които моделът да работи: information_gain, gain_ratio, accuracy, gini index и least_square. Дълбочината на дървото (броят на нивата в дървовидната структура) също може да се оптимизира, като се проверят всички възможни дълбочини от 1 до 100 със стъпка 10. Минималният брой обекти в листо също може да бъде тестван със стойности от 1 до 100 със стъпка 10. След задаването на възможните стойности на избраните параметри, методът се прилага като се проверяват възможните 1936 комбинации и от всички получени класификатори се предлага този с най-висока точност.

**Тълкуване на резултатите след подобряване на процеса**

Най-добре представилият се класификатор има следните параметри (Таблица 16):

*Таблица 12 Параметри на класификатора с най-висока точност, метод „Дърво на решенията“*

| **Наименование на параметрите в модела** | **Стойност** |
|---|---|
| Criterion | Gini Index |
| Maximal Depth | 0 |

| | |
|---|---|
| Minimal leaf size | 31 |
| Confidence | 0.05 |
| Split data – Sampling type | Linear sampling |

Оценката на най-добре представилият се класификатор след подобряването на процеса е представена на Фиг. 12, където се вижда, че цялостната точност на предсказване на получения класификатор (предсказване на двата класа) се е повишила до 79.53% (78.15%. преди подобряването), а точността на предсказване на класа на отпадащите клиенти също е значително повишена - 68.23% (59.96% преди подобряването).

**accuracy: 79.53%**

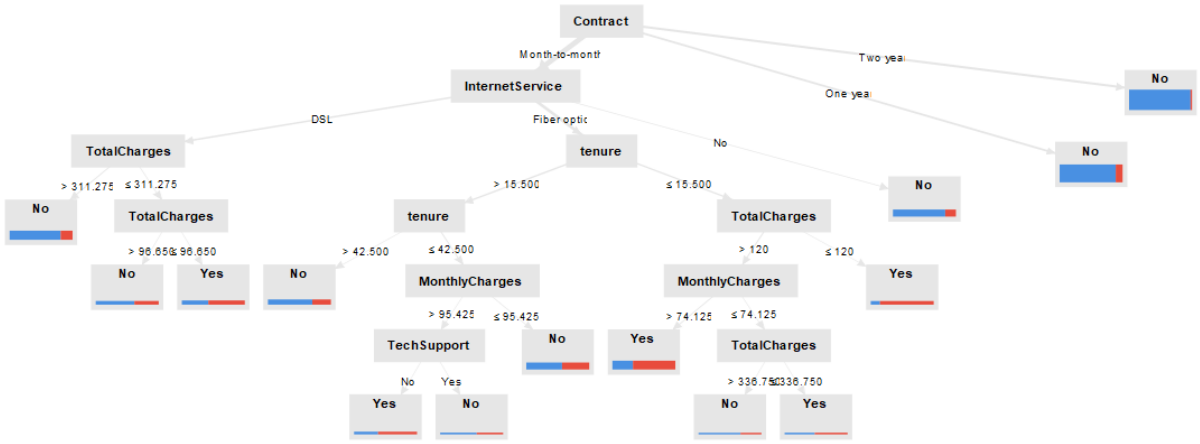
| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 1401 | 303 | 82.22% |
| pred. Yes | 129 | 277 | 68.23% |
| class recall | 91.57% | 47.76% | |

*Фигура 12 Матрица на класификацията получена след подобряване на модела при метода „Дърво на решенията“*

***Дърво на решенията и класификационни правила***

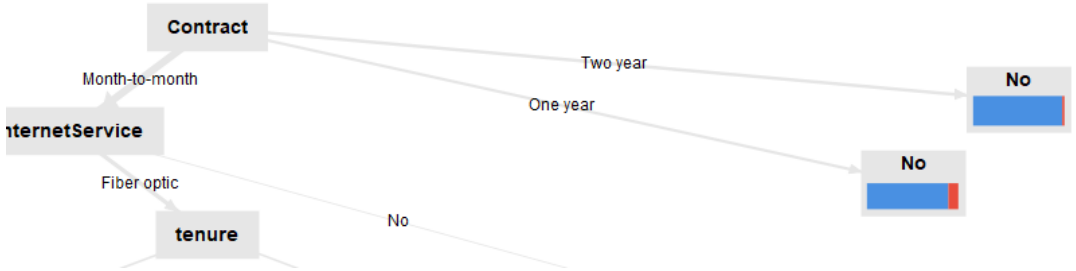
На фигура 13 е представена структурата на генерирания модел чрез метода „Дърво на решенията“. Избраният метод позволява лесно тълкуване, тъй като всяко разклонение е видимо в него. Коренът на дървото е представен от променливата „Contract“, с разделящ тест дали тези договори са за една година, две години или месец за месец. Разклоненията на второ ниво са свързани с променливата „Internet Service“. На третото ниво се срещат променливите „Total Charges“ и „Tenure“, а на четвърто ниво са видими - „Total Charges“ (два пъти) и „Tenure“. Петото ниво на разклоненията са представени от „Monthly Charges“ (два пъти). Последното разклонение, шест нива след корена на дървото, е осъществено с променливите „Total Charges“ и „Tech Support“.

Първото възможно разделяне на данните се извършва в корена на дървото, това означава, че според разделящия тест най-много на брой резултати могат да бъдат класифицирани в показани три групи правилно. Във всяко следващо разклонение попадат част от останалите данни, докато не са възможни повече деления. Колкото по-рано е използвана дадена променлива за деление, толкова по-голяма е нейната важност при разрешаване на зададения бизнес проблем.




*Фигура 13 Визуална презентация на полученият модел, чрез метод „Дърво на решенията“*

При първото деление, в корена на дървото, най-голям брой от клиентите са разпределени в Contract – Month to month (плащащи месечната си сметка без годишен или две-годишен договор) - 2727 записи или 55.40%, като за обучението на модела са използвани общо 4922 записа. Това означава, че променливата Contract дава възможност за най-добро първоначално класифициране на потребителите, склонни да прекратят договора си. При клиенти с две-годишен договор моделът показва 1146 клиента, които не са склонни да прекратят договора си и само 28, които са готови да го направят. Това са едва 2.38% от потребителите, попаднали в това разклонение. При клиенти с едногодишен договор моделът показва 910 човека от общо 1021, които не са склонни да прекратят договорите си, и 111, които могат да го направят. В този случай това са 10.87% от обектите, попаднали в това разклонение, отново това не е категоричен резултат, който да насочи усилията за задържане на клиентите в тази група от хора (фигура 14).

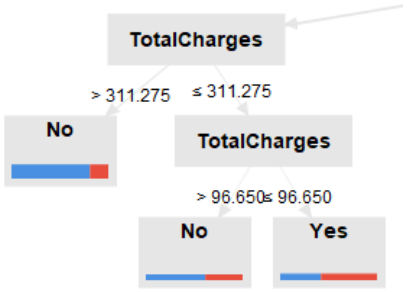



*Фигура 14 Първите листа на модела, получени при делението в корена на дървото на решенията*

Изводът от първото и най-важното деление на потребителите е, че не трябва фокусът да е върху клиенти с дългосрочни договори, защото тяхната склонност към прекратяване на

договора им е много ниска – общо 139 човека, потенциално отпадащи, от 2167 попаднали в тези две разклонения, което е едва 6,41% процент на напускащи клиенти.

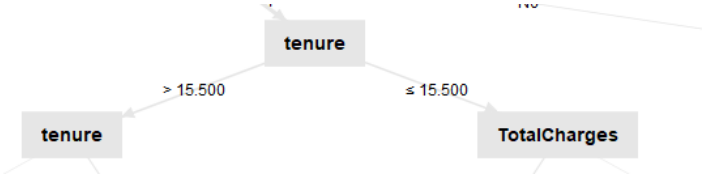
Следващият по важност тест е свързан с разклонението при променливата InternetService. В този случай 1501 (30.50%) от клиентите са класифицирани в клона ‘Fiber optic’. Другите две разклонения са съответно с 7.54% (InternetService = No) и 17.37% (InternetService = DLS) от общия брой записи, като в първата група преобладават ненапускащи клиенти (63 от 371 човека прекратяват своя договор или 16.98%, което не е достатъчно за категорични резултати за таргетиране на напускащи клиенти). В разклонението, където попадат 17.37% от потребители, се формира първата група (листо) където броят на напускащите клиенти е по-голям от ненапускащите (фигура 15.).




*Фигура 15 Първата група с преобладаващи напускащи клиенти*

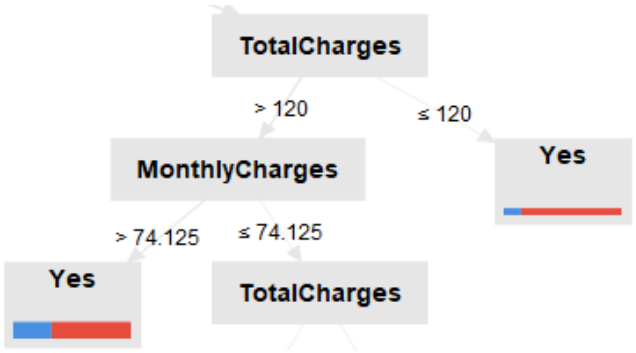
За да се стигне до това листо, трябва да бъдат направени още два теста: TotalCharges<=$311,275 и TotalCharges<=$96,650. В тази група попадат 197 човека, от които 114 са отбелязани като потенциални напускащи клиенти. Това са 58,76% потенциално отказващи се от услугите си обекти в групата.

Следващото разклонение е свързано с променливата „Tenure“ - продължителността от месеци, в които дадените потребители са клиенти на фирмата (фигура 16.). При този тест не се формират директно листа, а само други разклонения. Делението се случва при над 15,5 месеца и под 15,5 месеца продължителност на договора на потребителите, участващи в него.




*Фигура 16 Тест номер 3, след корена на „Дърво на решенията“*

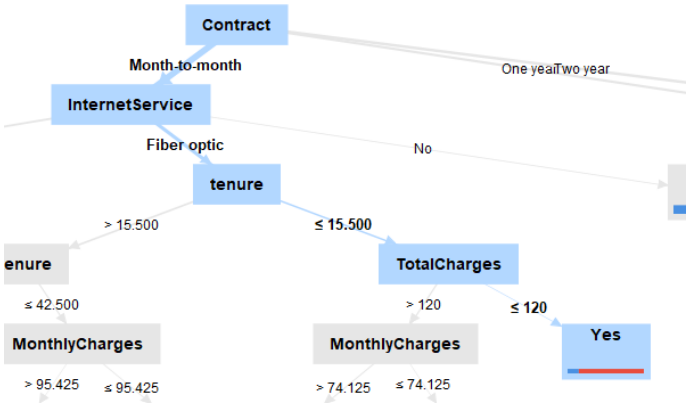
Достигането до следващите по-големи групи, в които напускащите клиенти са преобладаващи в листата на дървото на решенията, се случват на четвърто и пето ниво (фигура 17.).




*Фигура 17 Тестове на ниво 4 и 5, след корена на дървото*

Деленето на ниво 4 се формира посредством тестът „Total Charges“ - повече от $120 или по-малко и равно на $120. Полученото листо съдържа 163 записа, което е 3.31% от всички записи, използвани в модела. От тях 139 са маркирани като потенциално напускащи и 24 като ненапускащи клиенти, т.е. 85.27% от потребителите в групата потенциално ще прекратят своите договори.

Обобщение на извършените тестове, за да бъде достигнато това листо, е представено на фигура 18, подчертани в син цвят. Това представлява едно класификационно правило за клас „Yes“ (напускащи клиенти).



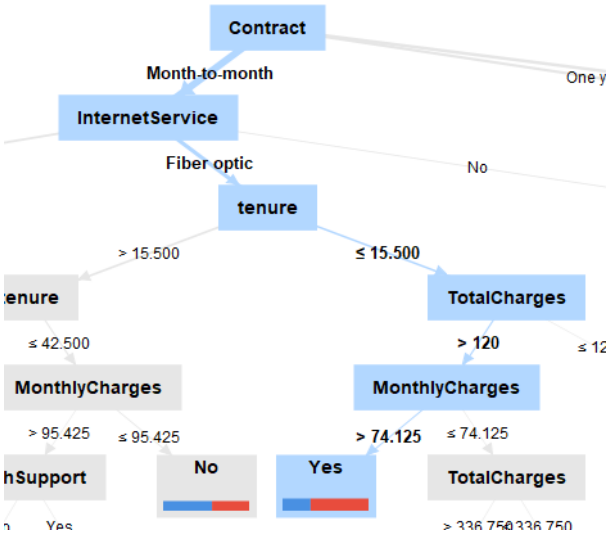

*Фигура 18 Достигане до листото с най-голям процент отпадащи клиенти, представено в четири последователни теста.*

Последователните тестове за достигане до това листо в дървото на решенията са както следва:

1. Contract = Month-to-month
2. InternetService = Fiber optic
3. Tenure <= 15 месеца и половина
4. Total Charges <= $120

Следователно всички потребители, които едновременно отговарят на посочените условия, е важно да бъдат внимателно проследени като потенциални отпадащи клиенти.

Последният тест от модела Дърво на решенията, който е детайлно разгледан, е представен на фигура 19, след делението MonthlyCharges > $74.125. Тази група (листо) е най-голямата (по брой записи) такава, в която преобладават потенциално прекратяващи своя договор клиенти. Групата съдържа 9.45% (475 броя) от записите, използвани за обучението на модела, като 320 потребителя са маркирани като склонни да напуснат фирмата. Успеваемостта на откриване на напускащи клиенти в това листо е 67.37%.




*Фигура 19 Достигане до листото с най-голям брой отпадащи, в което преобладават потребителите прекратили своя договор.*

Обобщение на направените тестове, подчертани в син цвят, е представено на фигура 19, и представлява още едно класификационно правило за клас „Yes“, съответстващ на отпадащите клиенти

Последователните тестове за достигане до това листо в дървото на решенията на пето ниво на разклоненията, са както следва:

1. Contract = Month-to-month
2. InternetService = Fiber optic
3. Tenure <= 15 месеца и половина
4. Total Charges > $120
5. MonthlyCharges > $74.125

Следователно всички клиенти, попадащи в тази група, е важно да бъдат внимателно анализирани и да се вземат необходимите мерки за тяхното задържане.

***Изводи:***

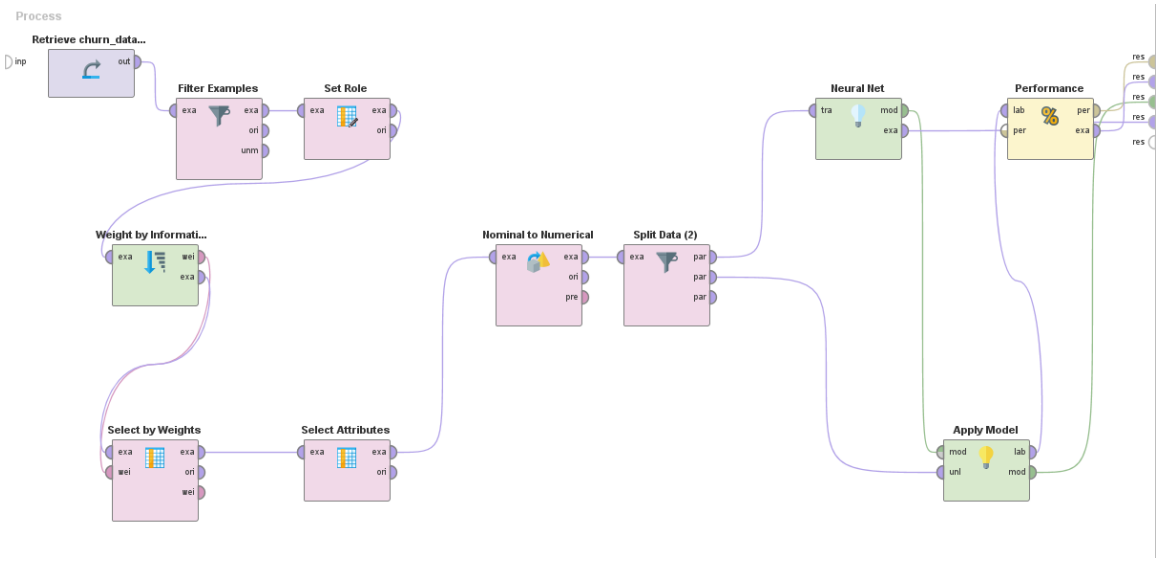
- Първоначалният класификатор, получен чрез метода „Дърво на решенията“, е с цялостна точност на предсказване (по отношение на двата класа) 78.15% преди

оптимизацията на модела, а след неговата оптимизация точността достига 79.53%. Предсказването на класа на отпадащите клиенти е подобрен от 59.96% на 68.23%, което доказва, че има смисъл от направената оптимизация на модела.

- Ниско рисковите потребители от това изследване имат общата характеристика, че техните договори са маркирани от променливата Contract като One Year и Two Year. Това са двете най-големи групи (листа) от модела с общо 2167 проверени записа, от които 2056 или 94.87% от клиенти не попадат в потенциално прекратяващите договора.
- Високо рискови групи, открити при това изследване, са клиенти със следните общи характеристики
    - Contract = Month-to-month
    - InternetService = Fiber optic
    - Tenure <= 15 месеца и половина
    - В тези групи попадат общо 638 потребители като 459 от тях потенциално биха прекратили договорите. Това са 71.94% от клиентите в тези две групи.
- На база на получените резултати от изследването може да бъде приложена стратегия, целяща високо рисковите потребители да подпишат срочен договор с продължителност една или две години, като по този начин част от тях ще останат за по-дълъг период от време във фирмата, без да прекратяват своите договори.

## 3.2. Генериране, оценка и подобряване на Data Mining класификационен модел с метода „Невронни мрежи“

Вторият генериран модел за откриване на напускащи клиенти в сферата на телекомуникациите, представен в тази разработка, е реализиран чрез метода „Невронни мрежи“. Цялостен изглед на процеса е представен на фигура 20.




*Фигура 20 Цялостен изглед на процеса изграден с метод „Невронни мрежи“, в инструмент RapidMiner Studio*

Основните стъпки в процеса по разработване на Data Mining модел за класификация чрез метода „Невронни мрежи“ са:

- Стъпка I – Зареждане на предварително избраните данни, като се избира техният източник (локален или достъпен в интернет пространство);
- Стъпка II – Филтриране на липсващи или грешни стойности;
- Стъпка III – Избор на променлива за предсказване;
- Стъпка IV – Избор на описателни променливи, които да участват в модела;
- Стъпка V – Преобразуване на категорийните данни в числови
- Стъпка VI – Разделяне на данните на две групи, за обучение и за тестване;
- Стъпка VII – Избор на Data Mining метода, който ще се използва за решаването на Data Mining задачата;
- Стъпка VIII – Прилагане на избрания Data Mining метода върху данните за обучаване;
- Стъпка IX – Оценява се точността на получения класификационен модел (класификатор) върху тестовите данни
- Стъпка X – Подобряване на процеса с цел получаването на класификатор с по-висока точност

До стъпка V процесът, изграден с метод “Невронни мрежи“, е идентичен с вече представения в точка 3.1 при метода „Дърво на решенията“. Първата основна разлика спрямо предходния представен процес се появява в стъпка V - Преобразуване на категорийни променливи в числови.

**V. Преобразуване на категорийни променливи в числови**

Преди на бъде избран Data Mining метод, който да бъде обучен и приложен върху данните, самите те трябва да бъдат подготвени за това. В тази част от дисертацията данните са анализирани с метода „Невронни мрежи“ и те трябва да бъдат подготвени в подходящ за това вид. За разлика от предходния процес, с метода „Дърво на решенията“, където алгоритъмът работи с категорийни данни, при метода „Невронни мрежи“ е необходимо всички променливи да са представени в числов вид.

Преобразуването на променливите от категорийни в числови е изпълнено с оператора, *nominal to numeric*. Параметърът coding type указва кодирането, което ще се използва за преобразуване на данните. Има три налични опции:

- unique_integers;
- dummy_coding;
- effect_coding.

**VI. Разделяне на данните на две групи, за обучение и за тестване**

При тази стъпка няма промяна в пропорциите на данните за обучение и тестване, описани при метода „Дърво на решенията“ - 70% от данни се използват за обучение, а останалите 30% за оценка на модела.

**VII. Избор на Data Mining метод**

В тази стъпка се определя Data Mining методът, който да бъде приложен върху обучаващата извадка от данни. При избор на метод „Невронни мрежи“ има възможност за избор на множество параметри като:

- Hidden Layers (Скрити слоеве): Този параметър съдържа името и големината/броя (hidden layer sizes) на всички скрити слоеве, в случая е създаден един скрит слой с два неврона;
- Training cycles (Тренировъчни цикли): Този параметър определя броя на циклите за обучение, използвани при прилагане на алгоритъм с Data Mining метода Невронни мрежи, в случая избраната стойност е 200;
- Learning rate (Скорост на обучение): Този параметър определя с колко се променят теглата на всяка итерация, в случая е избрана стойност 0.01;
- Momentum (Моментум): Този параметър добавя част от предишната актуализация на теглото към текущото тегло, за активиране на връзките между невроните.

**VIII. Прилагане на алгоритъма на избрания метод върху обучаващите данни**

За целта към оператора „Neural Net“ се подават обучаващите данни, които представляват 70% от общата съвкупност от данни. Стъпката е идентична с тази, показана по-рано при прилагане на метода „Дърво на решенията“.

## IX. Оценка на модела върху тестовите данни

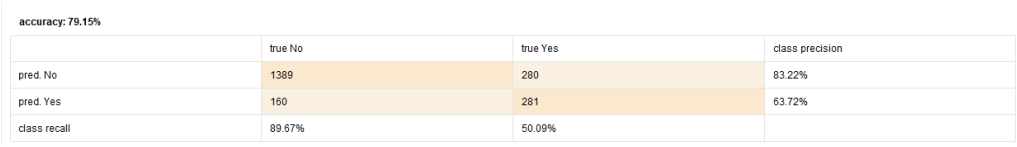
Последната стъпка е свързана с проверката на успеваемостта на модела, отново стъпката е сходна с вече поканата при метода „Дърво на решенията“. При създаване на процеса може да бъде зададен повече от един критерий за оценка на модела, по подразбиране е избрана опцията Accuracy, която представлява процентно съотношение на правилно предсказаните променливи спрямо всички налични. Другият избран критерии е матрица на класификацията (Confusion Matrix), който е част от споменатата по-горе опция, но показва точността на вярно и невярно предсказаните стойности на предсказваната променлива. Чрез този критерий може да се провери успеваемостта само на вярно предсказания клас на напускащите клиенти.

Първоначалният модел е получен  при следните зададени стойности на параметрите на алгоритъма (Таблица 17):

*Таблица 13 Първоначални настройки на алгоритъма, изпълнен с метод Невронни мрежи*

| Наименование на параметрите в модела | Стойност |
|---|---|
| Training cycles | 200 |
| Learning rate | 0.01 |
| Normalize | true |
| Hidden layers | 1 |
| Hidden layer size | 2 |

Оценката на цялостната точност на този класификатор (фигура 21) е 79.15% . Точността на предсказване на класа на отпадащите клиенти от матрицата на класификацията е 63.72%.



accuracy: 79.15%

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 1389 | 280 | 83.22% |
| pred. Yes | 160 | 281 | 63.72% |
| class recall | 89.67% | 50.09% | |

*Фигура 21 Резултат на модел за откриване на напускащи клиенти в сферата на телекомуникациите с метода „Невронни мрежи“*

## X. Подобряване на процеса

Подобряването на процеса е от ключово значение за постигането на по-високи резултати, които от своя страна ще допринесат за по-точно маркиране на потенциално напускащи клиенти и ще намалят финансовите загуби на фирмата. Оптимизацията на процеса е извършена по сходен начин на вече показния по-горе оператора „Optimize

Parameters (Grid)“. В него се поставя тази част от процеса, която се изпълнява след преобразуването на категорийните данни в числови.

*Таблица 14 Оптимизиране на настройките на алгоритъма, изпълнен с метод „Невронни мрежи“*

| **Наименование на параметрите в модела** | **Възможни стойности** |
|---|---|
| Training cycles | От 1 до 100, през 10 стъпки |
| Learning rate | От 0.01до 0.2 през 5 стъпки |
| Normalize | True / false |
| Split data – Sampling type | Automatic, shuffled, stratified и linear |

Извършва се експеримент като в оператора „Optimize Parameters“ се избират всички параметри за оптимизиране (Таблица 18). При избор на оператор „Neuron Network“ има възможност да бъдат избрани повече от един параметри, които да бъдат тествани в модела. В този случай са избрани: training_cycles, normalize, sampling_type, learning_rate. Броят на обучителни цикли може да се оптимизира като се проверят комбинации, вариращи от 1 до 100 опита със стъпка10. Може да бъде проверено дали нормализацията на данните оказва голямо значение върху резултатите на модела като се тества с параметър „normalize“. Важно е да се отбележи, че инструментът не позволява да се тества с различни стойности за броя скрити слоеве в структурата на невронната мрежа. След задаването на възможните стойности на избраните параметри, методът се прилага като се проверяват 528 комбинации и от всичките комбинации се предлага класификаторът с най-висока точност.

Новият модел е получен при следните настройки на параметрите на алгоритъма по метода „Невронни мрежи“ (Таблица 19):

*Таблица 15 Параметри на класификатора с най-висока точност, изпълнен с метод „Невронни мрежи“*

| **Наименование на параметрите в модела** | **Стойност** |
|---|---|
| Training cycles | 60 |
| Learning rate | 0.048 |
| Normalize | true |
| Hidden layer 1 | Size: 2 |
| Sampling type | automatic |

Получената цялостна точност на класификация (по отношение на двата класа) е 80.27% (фигура 22). Точността на предсказване за класа на отпадащите клиенти от матрицата на класификацията е 65.05%. Cohen’s kappa coefficient: 0.47.

accuracy: 80.27%

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 1841 | 331 | 84.76% |
| pred. Yes | 224 | 417 | 65.05% |
| class recall | 89.15% | 55.75% | |

*Фигура 22 Резултат от подобрен модел за откриване на напускащи в клиенти в сферата на телекомуникациите чрез метода „Невронни мрежи“.*

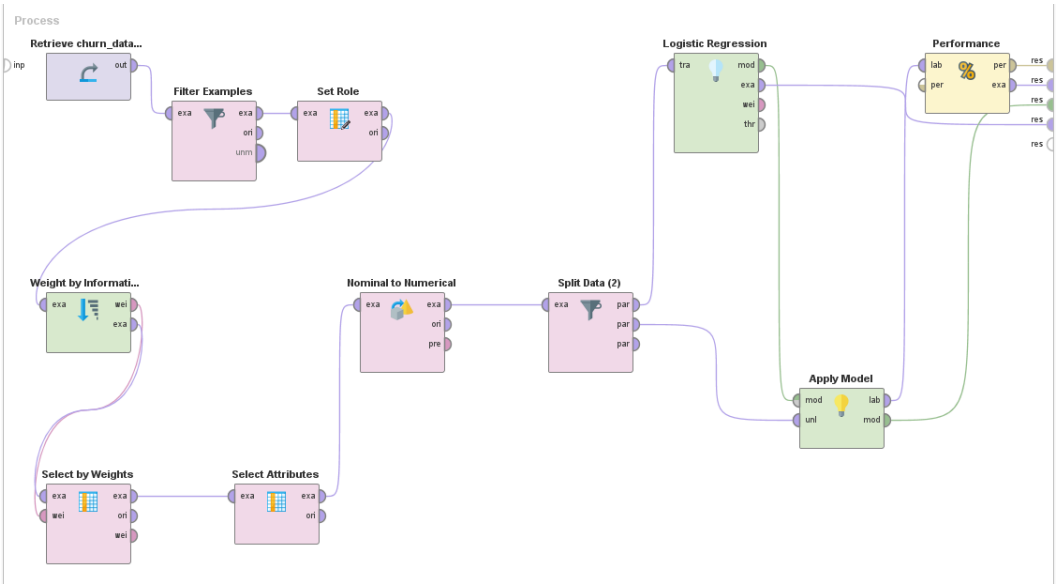
Характеристиките на новия модел са следните: един скрит слой с два неврона, активирането на невроните се случва посредством sigmoid алгоритъм (заложен от разработчиците на инструмента, не подлежи на промяна), позитивните тегла на променливите увеличават активните връзките между невроните, а негативните тегла на променливите намаляват активните връзки между невроните. В двата неврона най-висок коефициент има променливата Tenure. Колкото по-дълъг е договорът на клиента, толкова по-малко склонен е да прекрати своят договор – първи предсказван клас (Class ‘No’). Най-ниски коефициенти имат променливите „MonthlyCharges“ и „TotalCharges“ - те се свързват с втория предсказван клас (Class ‘Ýes’) – отпадащите клиенти.

***Изводи:***

- Резултатите от оценката на модела, получен чрез метода „Невронни мрежи“, са цялостна точност на предсказване (по отношение на двата класа) 79.15% преди оптимизацията, а след неговата оптимизация точността достига 80.27%. Точността на предсказването на класа на отпадащите клиенти е подобрено от 63.72% на 65.05%, което доказва, че има смисъл от направената оптимизация;
- Ниско рисковите потребители от това изследване имат общата характеристика, че техните договори, характеризирани от променливата „Tenure“, са с по-голяма продължителност като брой месеци.
- Високо рискови групи потребители, открити при това изследване, са обвързани с двете променливи „MonthlyCharges“ и „TotalCharges“ - колкото по-високи са стойности им, толкова по-склонни са потребителите да прекратят договорите си.

## 3.3. Генериране, оценка и подобряване на Data Mining класификационен модел с метод „Логистична регресия“

Третият генериран модел за откриване на напускащи клиенти в сферата на телекомуникациите, представен в тази разработка, е реализиран чрез метода „Логистична регресия“. Цялостен изглед на процеса е представен на фигура 23.




*Фигура 23 Цялостен изглед на процеса изграден с метод „Логистична регресия“, в инструмент RapidMiner Studio*

Основните стъпки в процеса при разработването на Data Mining модел за класификация чрез метода „Логистична регресия“ са:

- Стъпка I – Зареждане на предварително избраните данни, като се избира техният източник (локален или достъпен в интернет пространство);
- Стъпка II – Филтриране на липсващи или грешни стойности;
- Стъпка III – Избор на променлива за предсказване;
- Стъпка IV – Избор на описателни променливи, които да участват в модела;
- Стъпка V – Категорийните данни се преобразуват в числови
- Стъпка VI – Проверка за корелация между описателните променливи
- Стъпка VII – Разделяне на данните на две групи, за обучение и за тестване;
- Стъпка VIII - Избира се Data Mining метода, който ще се използва за решаване на Data Mining задачата за класификация;
- Стъпка IX – Алгоритъмът на избрания Data Mining метод се прилага върху данните за обучаване;
- Стъпка X – Точността на получения класификатор се оценява върху тестовите данни;
- Стъпка XI – Подобряване на процеса с цел получаване на по-добри резултати по отношение точността на предсказване.

До стъпка VI процесът, изграден чрез метод “Логистична регресия “, е идентичен с вече представения в точка 3.2. Първата основна разлика спрямо предходния представен процес се появява в стъпка VI. Проверка за корелация между описателните променливи променливите.

## VI. Проверка за корелация между описателните променливи

В тази стъпка от разработката на процеса е необходимо да бъдат проверени описателните променливите за корелация. Тази проверка е различна от направени тестове за избор на променливи в трета глава. По-рано беше обърнато внимание на корелацията между променливите при използването на линейни методи, какъвто е „Логистична регресия“. Важно е да бъде проверено дали има променливи с коефициент на корелация над 80%. Поради тази причина е използван оператора „Correlation Matrix“, чрез който коефициентите на корелация лесно могат да бъдат проверени (Фиг. 24).

| Attributes | InternetService | OnlineSecurity | OnlineBackup | DeviceProtection | TechSupport | Contract | tenure | MonthlyCharges | TotalCharges |
|---|---|---|---|---|---|---|---|---|---|
| InternetService | 1 | 0.607 | 0.650 | 0.663 | 0.609 | 0.100 | -0.030 | -0.322 | -0.176 |
| OnlineSecurity | 0.607 | 1 | 0.621 | 0.749 | 0.791 | 0.389 | 0.148 | -0.621 | -0.153 |
| OnlineBackup | 0.650 | 0.621 | 1 | 0.601 | 0.617 | 0.035 | -0.254 | -0.710 | -0.538 |
| DeviceProtection | 0.663 | 0.749 | 0.601 | 1 | 0.768 | 0.389 | 0.181 | -0.513 | -0.077 |
| TechSupport | 0.609 | 0.791 | 0.617 | 0.768 | 1 | 0.418 | 0.147 | -0.597 | -0.141 |
| Contract | 0.100 | 0.389 | 0.035 | 0.389 | 0.418 | 1 | 0.677 | -0.073 | 0.450 |
| tenure | -0.030 | 0.148 | -0.254 | 0.181 | 0.147 | 0.677 | 1 | 0.247 | 0.826 |
| MonthlyCharges | -0.322 | -0.621 | -0.710 | -0.513 | -0.597 | -0.073 | 0.247 | 1 | 0.651 |
| TotalCharges | -0.176 | -0.153 | -0.538 | -0.077 | -0.141 | 0.450 | 0.826 | 0.651 | 1 |

*Фигура 24 Матрица на корелацията*

От фигура 24 става ясно, че съществува корелация над 80% между две от използваните променливи – „tenure“ и „TotalCharges“. Поради тази причина е необходимо да бъде тествана точността на класификацията след премахване на всяка от тези две променливи.

## VII. Разделяне на данните на две групи, за обучение и за тестване

При тази стъпка няма промяна в пропорциите на данните за обучение и тестване, описани при методите „Дърво на решенията“ и „Невронни мрежи“ - 70% от данни се използват за обучение, а останалите 30% за оценка на модела.

## VIII. Избор на Data Mining метод

В тази стъпка се определя Data Mining методът, който да бъде приложен върху обучаващата извадка от данни. При избор на метод „Логистична регресия“ има възможност за избор на множество параметри като: Solver, Reproducible Use Standardize, Non-negative Coefficients, Add intercept, Compute p-values, Remove collinear columns, Missing values handling, Max interations, Max runtime

## IX. Прилагане на алгоритъма на избрания метод върху обучаващите данни

За целта към оператора „Logistic Regression“ се подават обучаващите данни, които представляват 70% от общата съвкупност от данни

Стъпката е идентична с тази, показана по-рано при прилагане на метода „Невронни мрежи“.

**X. Оценка на модела върху тестовите данни**

Последната стъпка е свързана с проверката на успеваемостта на модела, отново стъпката е сходна с вече поканата при метода „Невронни мрежи“. При създаване на процеса може да бъде зададен повече от един критерий за оценка на модела, по подразбиране е избрана опцията Accuracy, която представлява процентно съотношение на правилно предсказаните променливи спрямо всички налични. Другият избран критерии е матрица на класификацията (Confusion Matrix), който е част от споменатата по-горе опция, но показва точността на вярно и невярно предсказаните стойности на предсказваната променлива. Чрез този критерий може да се провери успеваемостта само на вярно предсказания клас на напускащите клиенти.

Първоначалният модел е получен  при следните зададени стойности на параметрите на алгоритъма (Таблица 20):

*Таблица 16 Първоначални настройки на алгоритъма, изпълнен с метод „Логистична регресия“*

| Наименование на параметрите в модела | Стойност |
|---|---|
| Solver | AUTO |
| Reproducible | False |
| Standardize | True |
| Non-negative coefficients | False |
| Add intercept | True |
| Compute p-values | True |
| Remove collinear columns | True |
| Missing values handling | Meanimplutation |
| Max iterations | 0 |
| Max runtime seconds | 0 |

Оценката на цялостната точност на този класификатор (фигура 25) е 78.91%. Точността на предсказване на класа на отпадащите клиенти от матрицата на класификацията е 63.06%.

accuracy: 78.91%

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 1385 | 281 | 83.13% |
| pred. Yes | 164 | 280 | 63.06% |
| class recall | 89.41% | 49.91% | |

*Фигура 25 Резултат на модел за откриване на напускащи клиенти в сферата на телекомуникациите с метода „Логистична регресия“*

## XI. Подобряване на процеса

Подобряването на процеса е от ключово значение за постигането на по-високи резултати, които от своя страна ще допринесат за по-точно маркиране на потенциално напускащи клиенти и ще намалят финансовите загуби на фирмата. Оптимизацията на процеса е извършена по сходен начин, като при предходните два метода, със следния оператор „Optimize Parameters (Grid)“. В него се поставя тази част от процеса, която се изпълнява след проверката за корелация между описателните променливи.

*Таблица 17 Оптимизиране на настройките на алгоритъма, изпълнен с метод „Логистична регресия“*

| **Наименование на параметрите в модела** | **Възможни стойности** |
|---|---|
| Solver | AUTO, IRLSM, L_BFGS |
| Split data – Sampling type | Automatic, shuffled, stratified и linear |
| Reproducible | True / False |
| Standardize | True / False |
| Non-negative coefficients | True / False |
| Add intercept | True / False |
| Compute p-values | True / False |
| Remove collinear columns | True |
| Missing values handling | Meanimplutation |
| Max iterations | 0 |
| Max runtime seconds | 0 |

Извършва се експеримент като от менюто „Operators“ се избрат всички параметри за оптимизиране (Таблица 21), без последните четири. Параметърът „Missing values handling“ остава по подразбиране, защото липсващи стойности не са допуснати до тази стъпка от изпълнението на алгоритъма. Параметрите „Max iterations“ и „Max runtime seconds“ остават със стойност 0, защото по този начин не са ограничени. Параметърът „Remove collinear columns“ трябва да остане избран като „True“, в противен случай алгоритъмът не може да произведе резултат при избор на стойност „True“ за параметър „Compute p-values“. Останалите параметри, които ще участват в оптимизацията, са видими в Таблица 3.9.

След задаването на възможните стойности на избраните параметри, методът се прилага като се симулират 384 класификатора и от всичките комбинации се предлага класификаторът с най-висока точност.

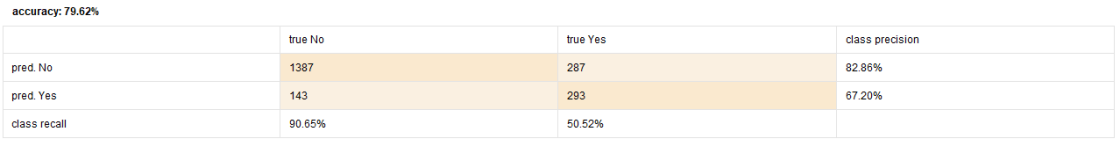
### Тълкуване на резултатите след подобряване на процеса

Новият, подобрен модел е получен при следните настройки на параметрите на алгоритъма по метода „Логистична Регресия“ (Таблица 22):

*Таблица 18 Параметри на класификатора с най-висока точност, метод „Логистична регресия“*

| Наименование на параметрите в модела | Стойност |
|---|---|
| Solver | IRLSM |
| Split data – Sampling type | Linear sampling |
| Reproducible | True |
| Standardize | True |
| Non-negative coefficients | True |
| Add intercept | True |
| Compute p-values | True |

Получената цялостна точност на класификация (по отношение на двата класа) е 79.62% (78.91% преди подобряването) (фигура 26). Точността на предсказване за класа на отпадащите клиенти, от матрицата на класификацията, е 67.20% (63.06% преди подобряването).



accuracy: 79.62%

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 1387 | 287 | 82.86% |
| pred. Yes | 143 | 293 | 67.20% |
| class recall | 90.65% | 50.52% | |

*Фигура 26 Резултат от подобрен модел за откриване на напускащи в клиенти в сферата на телекомуникациите чрез метода „Логистична регресия“*

Характеристиките на новия модел са следните: разделящ тест за определяне на класификационните групи: IRLSM. Разделянето на данните за обучение и данните за тестване е осъществено посредством метода: linear sampling. Всички останали параметри, за оптимизация, в модела (Reproducible, Standardize, Non-negative coefficients, Add intercept, Compute p-values) са със стойност: True.

Поради невъзможността да бъдат съчетани параметрите „Remove collinear columns“ и „Compute p-values“, е направен друг тест. Параметърът, избран за оптимизация, „Compute p-values“ е заменен с „Remove collinear columns“. Всички останали параметри са запазени от предходния тест.

Резултатите от тази промяна, след замяна на параметрите, са идентични с описаните на фигура 26.

След извършена проверка за корелация по-рано, бе установено, че има променливи с висок коефициент на корелация – „tenure“ и „TotalCharges“ (82%). При разработката на Data Mining модела за класификация, чрез метода „Логистична регресия“, е препоръчително да бъдат направени тестове с премахване на едната и после с другата от променливите в показаната двойка. Показаните тестове са направени с настройките, постигнали най-висока точност на предсказване на класа на отпадащите клиенти.

Резултатите след премахването на „TotalCharges“ (фиг 27):

**accuracy: 79.00%**

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 1386 | 299 | 82.26% |
| pred. Yes | 144 | 281 | 66.12% |
| class recall | 90.59% | 48.45% | |

*Фигура 27 Матрица на класификацията след премахване на променливата „TotalCharges“*

След премахване на променливата „TotalCharges“, която е първата от двете променливи с коефициент на корелация над 80%, точността на общия клас на предсказване, на оптимизирания модела, е спаднала от 79.62% на 79% и точността на класът на предсказване на отпадащите клиенти е намаляла от 67.20% на 66.12%.

Резултатите след премахването на „tenure“ (фиг 28):

**accuracy: 78.67%**

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 1390 | 291 | 82.69% |
| pred. Yes | 159 | 270 | 62.94% |
| class recall | 89.74% | 48.13% | |

*Фигура 28 Матрица на класификацията след премахване на променливата „tenure”*

След премахване на „tenure“, която е втората променлива от двойката с коефициент на корелация над 80%, точността на общия клас на предсказване, на оптимизирания модел, е спаднала от 79.62% на 78.67% и точността на класът на предсказване на отпадащите клиенти е намаляла от 67.20% на 62.94%.

Поради тези причини ще бъдат коментирани само описаните резултатите от първата оптимизация извършена върху алгоритъма, реализиран с метод „Логистична регресия“.

| Attribute | Coefficient | Std. Coefficient |
|---|---|---|
| InternetService | 0.550 | 0.406 |
| OnlineSecurity | -0.530 | -0.423 |
| OnlineBackup | 0.193 | 0.143 |
| DeviceProtection | -0.087 | -0.068 |
| TechSupport | -0.593 | -0.473 |
| Contract | -0.788 | -0.656 |
| tenure | -0.055 | -1.359 |
| MonthlyCharges | 0.009 | 0.283 |
| TotalCharges | 0.000 | 0.710 |
| Intercept | -0.537 | -1.771 |

*Фигура 29 Коефициенти и стандартизирани коефициенти на променливите участвали в алгоритъмът с най-висока точност, постигат чрез метод „Логистична регресия“*

Според модела получен, чрез метод „Логистична регресия“ променливите *InternetService* и *OnlineBackup* влият най-много на класифицирането на обектите. Това затвърждава резултатите свързани със значимостта на променливата *InternetService*, получени от моделът разработен с метод „Дърво на решенията“ в точка 3.1. Стандартизираният коефициент (фиг. 29) открива още две описателни променливи, които имат значение за крайните резултати от този модел. Това са „*MonthlyCharges*“ и „*TotalCharges*“, което отнотново е потвърдено от другите два метода, използвани в т.3.1 и т.3.2. Единствено променливата „tenure“, която вече е доказала важността си, в предходните изследвания, получава негативни стойности и за двата коефициента.

**Изводи:**

- Резултатите от оценката на модела, получен чрез метода „Логистична регресия“, са цялостна точност на предсказване (по отношение на двата класа) 78.91% преди оптимизацията, а след неговата оптимизация точността достига 79.62%. Точността на предсказването на класа на отпадащите клиенти е подобрено от 63.06% на 67.20%, което доказва, че има смисъл от направената оптимизация.
- Описателните променливи *InternetService, OnlineBackup,* „*MonthlyCharges*“ и „*TotalCharges*“ са определени като значими за правилното класифициране на обектите.
- Променливата „tenure“, която вече е доказала релеватността си, в моделите разработени с предходните два метода, за постигане на високи резултати на класификация, не е отбелязана като значима.

## 3.4. Сравнение на получените класификатори

Резултатите от оценката на класификаторите, получени чрез трите използвани метода в дисертационния труд, са представени в Таблица 24. Могат да бъдат направени следните изводи:

- Методът, с който са постигнати най-добри резултати при предсказване на класа на напускащите клиенти в това научно изследване, е „Дърво на решенията“. Получената точност на предсказване е 68.23%.
- Методът, с който са постигнати най-добри резултати при предсказване и на двата класа в това научно изследване, е „Невронни мрежи“. Получената цялостна точност на предсказване е 80.27%.
- Методът, с който е постигнат най-добър резултат според Cohen’s kappa coefficient, е „Невронни мрежи“, коефициентът е със стойност 0.47 (Таблица 24).

*Таблица 19 Сравнителна таблица за използваните методи по три избрани критерия: Обща успеваемост на предсказване, предсказване на класа на напускащите клиенти и Cohen’s kappa coefficient*

| **Метод** | **Точност на предсказване на двата класа** | **Точност на предсказване на класа на напускащите клиенти** | **Cohen’s kappa coefficient** |
|---|---|---|---|
| Дърво на решенията | 78.53% | 68.23% | 0.434 |
| Невронни мрежи | 80.27% | 65.05% | 0.47 |
| Логистична регресия | 79.62% | 67.20% | 0.446 |

Трите оптимизирани модела попадат в умереното ниво на съгласуваност, показващо възможността им да класифицират предсказания клас спрямо реалния.

Допълнително сравнение на класификаторите, получени чрез трите метода, е представено в Таблица 25, според критериите F1 – Score и AUC.

*Таблица 20 Сравнителна таблица за класификаторите, получени чрез трите метода, съгласно метриките F1 – Score и AUC*

| Метод | F1 - Score | AUC |
|---|---|---|
| Дърво на решенията | 56.19% | 0.819 |
| Невронни мрежи | 47.54% | 0.841 |
| Логистична регресия | 57.68% | 0.833 |

Стойността в колона F1 Score определя баланса между precision и recall (част от Матрица на класификацията за всеки от разработените модели). Колкото по близка до 1 е стойността в това сравнение, толкова по добър е балансът между false positive и false negative резултатите. Методите „Дърво на решенията“ и „Логистична регресия“, се представят с F1 Score над 51%, което означава приемливо ниво на баланс между false positive и false negative резултатите в тях, а методът „Невронни мрежи“ показва резултат 47.54%, което се свързва с недобър баланс на класовете. Трябва да се има предвид, че резултатите от тази метрика могат да варират според избрания метод. Добра стойност за модел, разработен с метод „Дърво на решенията“, не означа добра стойност за модел разработен с метод „Невронни мрежи“ (Buhl, 2023). Затова е избрана още една метрика, чрез която да бъде преценено, дали резултатите от разработените модели могат да бъдат счетени за релевантни.

Стойността в колоната AUC (Area Under the Curve / площта под ROC кривата) - измерва възможността на модела да различава класовете. Колкото по-близка до 1 е стойността на AUC, толкова по-добре алгоритъма поставя променливите в правилния клас. Според таблицата тълкуваща представянето на моделите спрямо AUC стойностите (в т. 1.4.3), моделите са се представили добре. AUC е над 0.81, което означава, че моделите се справят добре с различаването на класовете за предсказване.

## 3.5. Изводи по Трета Глава

В настоящата глава е постигнато следното:

- ✓ Обучени, оценени и подобрени са класификатори чрез избрани Data Mining методи за класификация, върху данни за клиенти на телекомуникационна компания, с Data Mining софтуер *RapidMiner Studio*.
- ✓ Обучените класификатори (получените Data Mining модели) са генерирани чрез избраните три метода за класификация – „Дърво на решенията“, „Невронни мрежи“ и „Логистична регресия“.
- ✓ За всеки от методите, процесите за генериране на класификаторите са оптимизирани, което във всеки от случаите води до подобряване както на цялостната точност на

класификаторите (по отношение предсказването на двата класа), така и на точността на предсказване на класа на отпадащите клиенти.

- ✓ След оптимизацията, най-висока точност на предсказване на класа на отпадащите клиенти се получава за класификатора, получен чрез метода „Дърво на решенията“ – 68.23%;
- ✓ При анализа на резултатите от оценяването на обучените класификатори са установени характеристиките на клиентите (описателните променливи), които най-добре определят напускащите клиенти - Tenure, MonthlyCharges, TotalCharges, Contract и Internet Service.
- ✓ Според използваните метрики за оценка Cohen’s Kappa coefficient и AUC и трите оптимизирани модела се справят добре при предсказването на класа на целевата променлива.

# 4. Заключение

Бизнес проблемът, свързан с навременното откриване на потенциалните напускащи клиенти в сферата на телекомуникациите, е изключително актуален и важен. Компаниите в телекомуникационния сектор полагат големи усилия в тази посока, тъй като е доказано, че разходите за придобиване на нови клиенти са в пъти по-високи от разходите за задържане на вече съществуващи клиенти, затова своевременното реагиране ще подпомогне намаляването на потенциалните парични загуби.

Основната цел на настоящата дисертация, да се генерират, изследват и подобрят Data Mining модели за класификация, като се разработи аналитичен модел за откриване на клиенти, прекратяващи своя договор (напускащи клиенти), е постигната.

На базата на подходящо избрана методика за реализация на Data Mining проекти, включваща CRISP-DM подход и софтуерен инструмент RapidMiner Studio, са обучени, оценени и подобрени класификатори чрез използването на три метода за класификация - „Дърво на решенията”, „Невронни мрежи“ и „Логистична регресия“, като най-висока точност на предсказване на класа на отпадащите клиенти се получава за класификатора, получен чрез метода „Дърво на решенията”.

При анализа на резултатите от оценяването на обучените класификатори са установени характеристиките на клиентите (описателните променливи), които най-добре определят рисковите клиенти. За отпадането на клиентите на телекомуникационните компании най-голямо значение имат типа на договора (Contract), продължителността на договорите на клиентите (tenure) и платените от тях сметки (MonthlyCharges, TotalCharges), типа на интернет услугата (Internet Service).

На базата на постиганите резултати можем да заключим, че поставената работна хипотеза в началото на дисертационния труд: “Възможно е обучените класификатори да

бъдат успешно използвани за предсказване класа на отпадащите клиенти на телекомуникационна компания“ е успешно доказана.

Апробираната методика може да бъде приложена върху нови различни данни, което да доведе до успешно идентифициране на отпадащи клиенти не само в сферата на телекомуникациите, но и в други области. Това са сфери, при които съществува подобен модел за месечен или годишен абонамент. Такива области са застраховане, облачни и хостинг услуги, стрийминг услуги като Netflix, Disney+, HBO и други.

Бъдещите изследвания могат да включват следните аспекти:

- Да бъдат открити допълнителни данни, свързани с потреблението на мегабайти и минути, на клиентите на компании от сферата на телекомуникациите. В използваната съвкупност от данни липсват променливи, съдържащи такава;
- Да се експериментира с промяна на пропорциите на данните за обучение и тестване на моделите, като е важно да се внимава с overfitting (прилепване/нагаждане) на модела към данните за обучение, при задаване на прекалено голям дял данни за обучение;
- Да се експериментира с архитектурата при метода „Невронни мрежи“, като се тества с различен брой скрити слоеве и различен брой неврони във всеки от слоевете. Подобни възможности не са налични през оператора за оптимизация в софтуерния инструмент RapidMiner Studio;
- Да се използват и други Data Mining методи за класификация, например „Random Forest“, „KNN“ и „SVM“, които не са сред най-често използваните, но също се срещат при решаването на бизнес проблема за откриване на отпадащи клиенти в телекомуникационния сектор.

# 5. Публикации по дисертационния труд

Извършените изследвания и получените резултати по темата на дисертационния труд са представени в изготвени четири публикации – 3 на английски език и 1 на български език, публикувани в сборници на международни конференции, проведени в България. Два от докладите са съвместна разработка на докторанта с неговия научен ръководител, а другите два са самостоятелни разработки.

Списък на публикациите по дисертационния труд:

1. Yanchev, H. (2019). Data Mining Methods Suitable for Customer Churn Discovery. *15th International Conference of ASECU*, (стр. 431–438). Sofia.
2. Yanchev, H. (2019). Data Mining методи подходящи за приложение върху неструктурирани данни от сферата на телекомуникациите. *Новите*

*информационни технологии и Големите данни: възможности и перспективи при анализите и управленските решения в бизнеса, икономиката и социалната сфера* (стр. 254-261). Sofia: UNWE.

3. Yanchev, H., & Kabakchieva, D. (2018). Data Mining Methods and Techniques used for Data Analysis in the Telecommunications Sector. *ICAICTSEE-2018*, (стр. 234-238). Sofia.
4. Yanchev, H., & Kabakchieva, D. (2024). Customer Churn Prediction in Telco Industry Using Artificial Neural Networks. *Innovative Information Technologies for Economy Digitalization – IITED* (стр. 323-332). Sofia: UNWE.

# Списък на таблиците

# Списък на фигурите

# Литература

Boujelbene, Z., Labidi, M. Y., Lemjid, A., Ncib, L., & Zghal, M. (2024). Customer churn prediction model enhacement for the telecommunication industry using data transformation methods and feature selection. *Global Scientific Journal vol. 12*, 2400-2417.

Buhl, N. (18 July 2023 r.). *F1 Score in Machine Learning*. Извлечено от Encord: https://encord.com/blog/f1-score-in-machine-learning/

Bujang MA, Sa’at N, Tg Abu Bakar Sidik TMI, Lim CJ. (2018). Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *Malays J Med Sci, 25*(4), 122–130.

Chowdhury, S., Lin, Y., Liaw, B., & Kerby, L. (2021). *Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance.* arXiv:2111.02513.

Chu, B.H., Tsai, M.S., Ho, C.S. (2007). Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems, 20*(8), 703-718.

Hadden, J., Tiwari, A., Roy, R., Ruta, D. (2007). Computer assisted customer churn management: state-of-the-art and future trends. *Computers & Operations Research, 34*(10), 2902-2917.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques Third Edition.*

Henry, F., Herwindiati, D., Mulyono, S., & Hendryli, J. (27 May 2016 r.). Sugarcane Land Classification with Satellite Imagery using Logistic Regression Model. *International Conference on Information Technology and Digital Applications.* Извлечено от https://www.droracle.ai/articles/137703/what-is-the-interpretation-of-kappa-kappa-statistic-agreement

Kaggle. (2016). *kaggle.com.* Извлечено от kaggle.com: https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data

KhakAbi, S., Gholamian, M.R., & Namvar, M. (2010). Data Mining Applications in Customer Churn Management. *International Conference on Intelligent Systems, Modelling and Simulation,* 220-225.

Khalid, L. F., Abdulazeez, A. M., Zeebaree, D. Q., Ahmed, F. Y., & Asaad, D. (2021). Customer Churn Prediction in Telecommunications. *2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, (стр. 81-86).

Khan, M. R., Manoj, J., Singh, A., & Blumenstock, J. (2015). Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty. *2015 IEEE International Congress on Big Data*. doi:10.1109/bigdatacongress.2015.107

Lazarov, V., & Capota, M. (2007). *Churn Prediction.* Technische Universität München.

Mamcenko, J., & Gasimov, J. (2014). Customer Churn Prediction in Mobile Operator Using Combined Model. *ICEIS*.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., & Johnson, E. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks, 11*(3), 690 – 696.

Ng, K., & Liu, H. (2000). Customer Retention via Data Mining. *Artificial Intelligence Review, 14*(6).

Nurtriana, A. (2024). Churn prediction analysis of telecom customers using svm, random forest and logistic regression models using orange data mining tools. *ICCSEI 2023.*

Owczarczuk, M. (2010). Churn Models for Prepaid Customers in the Cellular Telecommunication Industry Using Large Data Marts. *Expert Systems with Applications, 37*(6), 4710-4712. doi:https://doi.org/10.1016/j.eswa.2009.11.083

Poudel, S. S., Pokharel, S., & Timilsina, M. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. *Machine Learning with Applications vol. 17*.

Rusu, O. & Halcu, Ionela & Grigoriu, O. & Neculoiu, Giorgian & Sandulescu, Virginia & Marinescu, M. & Marinescu, Viorel. (2013). *Converting unstructured and semi-structured data into knowledge.*

Sadiku, M. N., Adekunte, P. A., & Sadiku, J. O. (2024). Big Data in Telecommunications. *International Journal of Trend in Scientific Research and Development (ijtsrd)*, 243-252.

Saini, N., & Garg, D. K. (2017). Churn Prediction in Telecommunication Industry using Decision Tree. *International Journal of Engineering Research & Technology (IJERT)*, (стр. 439-443).

Saleh, M. E., & Abd-Alsabour, N. (2024). Improved Decision Tree, Random Forest, and XGBoost Algorithms for Predicting Client Churn in the Telecommunications Industry. *International Journal of Advanced Computer Science and Applications, vol. 15*, 674-682.

Stefanova, K., & Kabakchieva, D. (2015). *Business Intelligent Systems.* University of National and World Economy.

Terra, J. (31 July 2025 r.). *What is a ROC Curve, and How Do You Use It in Performance Modeling?* Извлечено от simplilearn: https://www.simplilearn.com/what-is-a-roc-curve-and-how-to-use-it-in-performance-modeling-article

Tsai, C.F., Lu, Y.H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications, 36*(10), 12547-12553.

Tuğba, U., Gürsoy, Ş. (2010). Customer Churn Analysis in Telecommunication Sector. *Journal of the School of Business Administration, 39*(1), 35-49.

Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in Regression Analyses Conducted in. *Epidemiology (Sunnyvale).*

Vuval, U. (2020). Churn Prediction for Telecommunication Industry Using Artificial Neural Networks. *International Journal of Computer and Information Engineering,* 396-399.

Wei, Y. (2024). Telco Customer Churn Prediction. *Highlights in Science, Engineering and Technology vol.92*, (стр. 218-226).

Weiss, G. M. (2009). *Data Mining in the Telecommunications Industry.* Fordham University.

Witten, I. H. & Frank, E. (2005). *Data Mining Practical Machine Learning Tools and Techniques (2nd ed.).* San Francisco, CA: Morgan Kaufmann Publishers Inc.