Технології Big Data у Аналізі Ризиків Страхової Компанії

Михайло Лазар
кафедра програмної інженерії
Харківський національний університет
радіоелектроніки
Харків, Україна
mykhailo.lazar@nure.ua

Володимир Кобзєв
кафедра програмної інженерії
Харківський національний університет
радіоелектроніки
Харків, Україна
volodymyr.kobziev@nure.ua

Big Data Technologies in the Insurance Risk Analysis

Mykhailo Lazar
Software Engineering department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
mykhailo.lazar@nure.ua

Volodymyr Kobziev
Software Engineering department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
volodymyr.kobziev@nure.ua

***Анотація*—Розглядається можливість використання технологій Big Data у страхуванні. Зокрема, розглядається можливість використання сучасних засобів збору характеристик клієнтів, які використовуються в аналізі страхових ризиків у клієнтоорієнтованих страхових продуктах.**

***Abstract*—The possibility of using Big Data technologies in insurance is considered. In particular, the possibility of using modern means of collecting customer characteristics used in the analysis of insurance risks in client-oriented insurance products is considered.**

Ключові слова— страхування; страхові ризики; оцінки ризику; Big Data; кластеризація

Keywords— insurance; insurance risks; risks estimates; Big Data; clasterization

VIII. Вступ

З появою теорії ймовірностей і актуарної науки як математичних дисциплін в XVII столітті, аналіз даних зіграв фундаментальну роль в страхуванні. Ці наукові досягнення дозволили страхуванню розвитися в галузь, засновану на раціональному обчисленні наявних даних і прийнятті обґрунтованих рішень. Ключову роль в зборі даних, необхідних для розрахунку актуарного аналізу зіграла Церква. У XVI столітті в деяких європейських країнах парафіяльним священикам було наказано вести записи про кількість хрещень, шлюбів і поховань [1].

У 1693 році Едмунд Галлей побудував першу повну таблицю смертності для населення міста Бреславля (Вроцлав), включивши в неї дитячу смертність. Він дав визначення основних показників таблиці смертності, перелічив ймовірності дожиття і смерті для своїх сучасників, ввів в науку поняття середньої тривалості майбутнього життя, сформулював методику регулювання тарифів в страхуванні життя за допомогою таблиці смертності. Фактично, Галлей є засновником теорії актуарних розрахунків у сфері страхування життя. Він ввів поняття норми відсотка або норми зростання грошей в страхуванні. Форма таблиці смертності Галлея і принципи її побудови використовуються в страхуванні донині [2].

В минулому страховики в основному покладалися на особисту інформацію зібрану безпосередньо у клієнтів (страхувальників) під час андеррайтингу для угруповання осіб в класи ризику. Наприклад, при особистому автострахуванні, страховики зазвичай покладалися на відомості про тип і вік автомобіля, історію втрат. За останні два десятиліття страховики все частіше використовують додаткові дані зі сторонніх джерел даних. Наприклад, коли з'явилися емпіричні дані про те, що люди з більш високими кредитними балами також мають тенденцію бути більш обережними водіями, страховики почали включати кредитні бали в їх аналіз для особистого автострахування. Роль даних, однак, в основному залишилася колишньою, їх використовують для розуміння ризиків і компенсації збитків, завданих клієнтам.

IX. BIG DATA ТА СТРАХУВАННЯ

Сьогодні прогрес в області Big Data, штучного інтелекту та «Інтернету речей» (IoT) приводять до фундаментальної трансформації ролі даних в страховій бізнес-моделі. Цей розвиток викликаний появою двох нових джерел даних, які є актуальними в контексті страхування. Перший складається з даних, які автоматично генеруються в онлайн-режимі. Такі дані містять особисту інформацію, надану через соціальні мережі, через сервіси онлайн-покупок, а також дані, отримані у результаті пошуку та перегляду інформації. Дані про поведінку в Інтернеті допомагають виявляти інформацію про звички та спосіб життя людей доповнюють або змінюють дані, що традиційно використовуються страховими компаніями. Друге нове джерело даних пов'язане з вбудованими датчиками в побутові прилади та інші споживчі товари в IoT, наприклад датчиків, вбудованих в автомобілі або дані з носимих пристроїв, дані з розумних будинків у яких певний час знаходяться клієнти. Ці дані, зазвичай, фрагментовані і специфічні для конкретної мети.

Безперервний збір і аналіз поведінкових даних дозволяє робити індивідуальну і динамічну оцінку ризиків і встановлення безперервного циклу зворотного зв'язку для клієнтів, без втручання або обмеженого втручання людини. Такий цифровий моніторинг не тільки підвищує якість оцінки ризику, але також може надавати інформацію в режимі реального часу страхувальникам за їх ризикової поведінки та слугувати індивідуальним стимули для зниження ризику. Поєднання нових джерел даних - шлях для впровадження розширеного управління ризиками. Систем, які використовують інтелектуальну аналітику як основу для раннього втручання і запобіганню ризиків. Такі нові бізнес-моделі вже запущені або вже на горизонті. Вони включають як компанії, які пропонують індивідуальний страховий продукт, так і повністю цифрових страховиків (наприклад, Oskar, In Shared, Haven Life або Sherpa) [1].

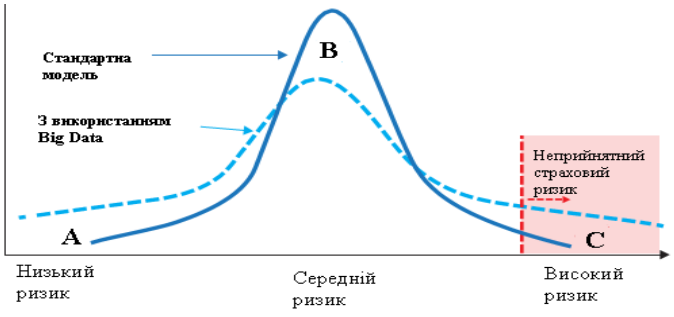

Рис. 1 Розподіл страхових ризиків

Посилення аналізу даних є вигідним як для страхувальників, так і їх клієнтів. Використання технологій Big Data дозволить страховикам краще розрізняти ризики. Цей процес проілюстровано на діаграмі (рисунок 1). Завдяки більш точному розподілу частка застрахованих як «Середній ризик» (B) буде зменшуватися вони будуть перекваліфіковані у «Низький ризик» (A) або «Високий ризик» (C). Таким чином у страхової компанії з'являється можливість зменшити розмір страхових премій для застрахованих з низьким ризиком і навпаки збільшити премії для застрахованих з високим ризиком. Такий перерозподіл розміру страхових премій дозволить знизити навантаження на тих клієнтів де ризик є низьким та більш ефективно виконувати свої зобов'язання перед клієнтами з високим ризиком. Тож використання технологій Big Data у страхуванні дозволяють клієнтам платити саме за той рівень захисту, який вони потребують, а страховим компаніям зробити свій ризик-менеджмент адекватним наявним ризикам.

Вже існує програма телематики «Progressive Insurance» (США), яка включає в себе продукт «Pay as You Drive» - пристрій встановлений в автомобілі застрахованого водія. Він збирає дані, такі як час дня, швидкість руху транспортного засобу і тенденції гальмування [3].

Звісно встановлення спеціального пристрою в кожен застрахований транспортний засіб пов’язане з певними витратами. Цю проблему можна вирішити шляхом встановлення спеціального програмного забезпечення на смартфон водія, що буде використовувати наявний GPS пристрій для визначення місцезнаходження транспортного засобу. Таким чином за допомогою мобільних додатків у смартфонах, смарт-годинників чи фітнес-трекерів можна збирати необхідну інформацію не тільки про транспортні засоби але і про людину. Такі данні є дуже важливими я в маркетингових цілях так і для страхування життя і здоров'я.

Використання технологій Big Data у сфері страхування здоров'я дозволить не тільки визначати ризик захворювання чи смерті для кожної людини окремо але і попередити їх. Маючи можливість одночасно обробляти інформацію про конкретну людину та накопичені системою охорони здоров’я статистичні дані можна з великою точністю передбачити ту чи іншу хворобу чи навіть її попередити. Смартфон чи інший носимий пристрій за допомогою відповідного програмного забезпечення зможе надавати певні рекомендації чи повідомити про необхідність відвідати лікаря. Страхова компанія може оптимізувати свої ризики, а клієнт отримати саме той рівень захисту який потрібен конкретно у його випадку.

Дослідження показують, що програми на основі способу життя, пов'язані з використанням носіїв, таких як «smart watches», стають все більш популярними. Ці програми впроваджують страховики, які збирають особисті дані окремих осіб, відстежують поведінку застрахованої особи, включаючи кількість виконаних кроків, кількість хвилин активності протягом дня, відпочинку і серцевого ритму і якість сну. Вони

призначені для позитивного впливу на поведінку страхувальника шляхом встановлення цілей і стимулів. Це також дає більш оптимальні страхові внески, які більш точно відображають ризик страхувальника. Дослідження показує [3], що 93% роздрібних клієнтів (в Австралії, Франції, Німеччині, Великій Британії та США) готові поділитися персоналізованими даними, якщо вони можуть заощадити гроші або отримати індивідуальні пропозиції.

Революція IoT та розвиток таких систем, як «Розумний будинок», відкривають нові можливості для оптимізації страхових ризиків і у напрямку страхування нерухомого майна. Головною метою таких систем є автоматизація будинку та побутової техніки. Але ці системи можна використовувати і для покращення захисту нерухомості. У такому «Розумному будинку» можна слідкувати за вологістю повітря, температурою, сейсмічною активністю, навантаження на електропроводку тощо. На підставі аналізу цієї інформації у разі необхідності автоматично будуть виконуватися певні дії (включення протипожежних систем, виклик аварійних служб та ін.). Також ці дані окрім попередження небажаних подій дозволять чітко визначити ризики, притаманні тому чи іншому об'єкту. Так, при страхуванні буде прийматися до уваги обладнаний об'єкт страхування такими системами чи ні, і в залежності від цього страхова премія буде збільшуватися або зменшуватися.

Звісно окрім переваг постійного збору даних виникає питання, як ця інформація буде використовуватися і чи не є це втручанням в особисте життя. У цьому питанні дійсно важливо саме як і для чого використовується зібрана інформація. Адже для того щоб страховій компанії правильно визначити ризики настання тої чи іншої події у конкретному випадку зовсім не потрібно слідкувати за кожним кроком людини. У тому ж прикладі з автомобілем нема потреби постійно слідкувати за його пересуванням. Достатньо мати дані про основні маршрути пересування. Маючі цю інформацію можна розрахувати ймовірність настання страхового випадку, використовуючи статистичні дані саме на цих маршрутах, а не в середньому по місту регіону чи країні. І ці ризики можуть суттєво відрізнятися впливаючи як на рівень захисту страхувальника так і на рівень страхової премії.

З метою групування отриманих даних за певними ознаками доцільно використовувати алгоритми ієрархічного поділу (partitioning algorithms). Ці алгоритми здійснюють декомпозицію набору даних, що складається з n спостережень, на k груп (кластерів) із заздалегідь невідомими параметрами. При цьому виконується пошук центроїдів - максимально віддалених один від одного центрів згущення точок C_k з мінімальним розкидом всередині кожного кластера [4]. Розглянемо приклад використання методу k середніх Мак-Кіна (k-means clustering, MacQueen, 1967), в якому кожен з k кластерів представлений центроїдом.

Страхова компанія «ABC» періодично купує списки водіїв із зовнішніх джерел. Актуарії в компанії «ABC» хочуть оцінити частоту потенційних претензій для цілей андеррайтинга. Для цього необхідно розділити водіїв, які є подібні один до одного, на групи відповідно до потенційного ризику. Після того, як водії будуть сегментовані, розмір вибірки потенційних клієнтів у кожному сегменті може бути використаний для оцінки частоти претензій. Результати цієї тестової оцінки дозволять актуаріям оцінити потенційний прибуток від потенційних клієнтів зі списку, як у цілому, так і для конкретних сегментів. Опис даних, які отримуються від постачальника, наведені в таблиці 1.

ТАБЛИЦЯ 1. ДАНІ ВОДІЇВ АВТОМОБІЛІВ

Змінна	Тип змінної	Рівень вимірювання	Опис
Age	Неперервний	Інтервальний	Вік водія в роках
Car age	Неперервний	Інтервальний	Вік автомобіля в роках
Car type	Категоричний	Номінальний	Тип автомобіля
Gender	Категоричний	Бінарний	F = жінка, M = чоловік
Coverage level	Категоричний	Номінальний	Політика охоплення
Education	Категоричний	Номінальний	Рівень освіти
Location	Категоричний	Номінальний	Місце проживання
Climate	Категоричний	Номінальний	Клімат-код
Credit rating	Неперервний	Інтервальний	Кредитна оцінка водія
ID	Вхідний	Номінальний	Ідентифікаційний номер водія

Для формування кластерів найчастіше використовують метод K-means. Розрахунки можна зробити за допомогою бібліотеки «cluster» мови R.

У таблиці 2 наведено принт-скрин результату кластеризації.

Відхилення означає середньоквадратичну помилку серед змінних кластерного стандарту відхилення, що дорівнює середньоквадратичній відстані між випадками в кластері. Під час процесу кластеризації значення обчислюється як значення між 0 та 1 для кожної змінної. Значення є мірою вартості заданої змінної для формування кластері. Як показано в таблиці 3, змінна "Gender " має значення 0, це означає, що ця змінна не використовувалась як розподіляюча змінна при розробці кластерів. Міра "Importance" вказує на те, наскільки добре кожна змінна розподіляє дані на класи. Змінні з нульовим значення не повинні бути виключені до процесу класифікації.

ТАБЛИЦЯ 2. ХАРАКТЕРИСТИКИ КЛАСТЕРІВ

Cluster	Frequency of Cluster	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Credit Score	Car Age	Age	Gender	Location	Climate	Car Type	Coverage	Education
9	7	2.87	5	2.82	0.86	3.29	35.57	1.00	3.43	1.29	3.57	2.43	1.86
8	20	3.22	7	2.40	0.62	2.15	46.65	0.65	2.80	2.55	2.25	2.85	1.85
7	22	3.25	2	2.25	0.65	2.73	24.59	0.27	1.95	2.09	1.45	2.36	2.27
6	21	3.38	4	2.41	0.81	6.52	35.19	0.43	2.00	1.48	1.67	1.19	1.76
5	33	3.41	4	2.37	0.82	3.00	32.79	0.58	3.82	2.33	2.03	2.39	3.03
4	18	3.83	5	2.37	0.59	5.17	34.44	0.39	3.50	1.83	2.72	1.44	2.56
3	7	3.21	7	3.14	0.46	8.00	20.57	0.43	3.57	2.43	1.14	1.43	2.00
2	18	3.38	7	2.25	0.56	3.56	26.00	0.28	2.89	2.67	1.28	2.28	1.39
1	27	3.40	5	2.55	0.75	2.37	44.15	0.07	2.04	1.52	3.30	2.70	3.00

ТАБЛИЦЯ 3. ВАЖЛИВІСТЬ ЗМІННИХ

Name	Importance
GENDER	0
ID	0
LOCATION	0
CLIMATE	0
CAR_TYPE	0.529939
COVERAGE	0.363972
CREDIT_SCORE	0.343488
CAR_AGE	0.941952
AGE	1
EDUCATION	0.751203

Кластерний аналіз може бути використаний для покращення страхування від нещасних випадків шляхом сегментування баз даних у більш однорідні групи ризику. Однорідну групу можна досліджувати, аналізувати та моделювати. Сегменти за типом змінних, що асоціюються з факторами ризику, прибутком або поведінкою, часто забезпечують різкі контрасти, які буде легше тлумачити. Як результат, актуарії можуть більш точно передбачити ймовірність вимог та суму позову [5].

Нарешті, ідентифікатор кластера для кожного спостереження може бути переданий іншим вузлам для використання як вхідної, ідентифікаційної, групової або цільової змінної. Наприклад, можна сформувати кластери на основі різних вікових груп, на які націлений страховий агент. Потім можна створити прогнозні моделі для кожної вікової групи, передаючи змінну кластера як групову змінну до моделюючого вузла [6].

X. ВИСНОВКИ

Використання технологій Big Data відкриває великі перспективи використання у всіх сферах людського життя і в страхуванні зокрема. Використання великої кількості додаткових характеристик клієнтів дозволить ще точніше враховувати ризики та за рахунок використання більш адекватної моделі страхування забезпечити бажаний рівень діяльності страхової компанії.

ЛІТЕРАТУРА REFERENCES

[1] Big Data and Insurance:Implications for Innovation, Competition and Privacy March 2018.

[2] Колчинский И.Г., Корсунь А.А., Родригес М.Г. Астрономы: Биографический справочник. — 2-е изд., перераб. и доп.. — Киев: Наукова думка, 1986. — 512 с.

[3] The Impact of Big Data on the Future of Insurance Green Parer November 2016

[4] В.К. Шитиков, С.Э. Мастицкий “Классификация, регрессия и другие алгоритмы Data Mining с использованием R”- Тольятти, Лондон – 2017.

[5] “Applying Data Mining Techniquesin Property/Casualty Insurance” Lijia Guo, Ph.D., A.S.A. University of Central Florida.

[6] A.B. Devale and Dr.R.V. Kulkarni “Applying Data Mining Techniquesin inlife insurance” - International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.4, July 2012.