DOI: 10.31489/2023Ec1/152-161
JEL D58, C58, I13
ӘОЖ 330.47

Н.Н. Гелашвили[1*], А. Сабыржан[2], Б.Х. Раимбеков[3], Ғ.Ә. Кенешева[4], Г.Қ. Абдраманова[5].

[1,2,3] Е.А. Бөкетов атындағы Қарағанды университеті, Қарағанды, Қазақстан;
[4]Қазтұтыну одағы Қарағанды университеті, Қарағанды, Қазақстан;
[5]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

[1]denor1980@mail.ru, [2]alisher-aliev-79@mail.ru, [3]rbh2006@yandex.kz, gizzat@yandex.ru,[5]agk2009@mail.ru

[1]https://orcid.org/0000-0002-7115-2007, [2]https://orcid.org/0000-0003-4619-9951, [3]https://orcid.org/0000-0001-5288-6059, [4]https://orcid.org/0000-0003-2928-8928, [5]https://orcid.org/0000-0001-5268-8706

[1]Scopus Author ID:57215350667, [2]Scopus Author ID:57222593357, [3]Scopus Author ID:57190606419, [4]Scopus Author ID:57208010004, [5]Scopus Author ID:57195629019

Медициналық сақтандыру жүйесінде деректерді өндіру әдістерін практикалық қолдану

Аңдатпа:

Қазіргі кезеңде ақпарат компанияның іскерлік қызметінің сәттілігінің маңызды факторына айналды. Компанияның бәсекегеқабілеттілік деңгейі оның болуына, сапасына және мерзіміне тығыз байланысты. Сонымен қатар, бүгінде келіп түскен ақпараттың саны өте көп және оны тиімді өңдеу қажеттілігі туындайды.

Зерттеудің мақсаты болашақ денсаулық сақтандыру төлемдерін болжау үшін *Data Mining* құралын пайдаланып ақпаратты өңдеудің математикалық әдістерін қолдану болып табылады.

Әдісі: Болашақ сақтандыру төлемдерінің мөлшерін болжау және сақтандыру жағдайларының пайда болу факторларын анықтау үшін сызықтық және көпмүшелік регрессия әдістері таңдалған. Алгоритмдер көрсеткішінің негізгі факторлары ретінде: клиенттердің жынысы мен жасы, темекі шегу әдетінің болуы, дене салмағының бойға қатынасы, тұрғылықты жері, сақтандыруы бар балалардың болуы, медициналық шығындардың мөлшері және тұрғылықты жері пайдаланылды.

Нәтижелер: Жасы мен темекі шегу денсаулыққа қатысты сақтандыру жағдайының деңгейіне үлкен мән береді. Бұл сақтандыру компаниясының клиенттерінің әл-ауқатына ең үлкен теріс әсер етуі мүмкін. Дене салмағының индексі де үлкен мәнге ие. Ең азы денсаулық сапасына тұрғылықты жері мен жасы әсер етеді. Мұның бәрі сақтандыру компаниясының басшылығына сақтандыру өтемақыларын төлеу шығындарын нақты жоспарлауға мүмкіндік береді. Сондай-ақ, сақтандыру жағдайының ықтималдығына әсер ету дәрежесін анықтау сақтандыру компаниясына клиентке келтірілген залалды өтеу тетіктері сипатталатын шарттарды неғұрлым икемді жасауға мүмкіндік береді.

Қорытынды: *Data Mining* компания қызметін тиімді басқарудың тиімді құралы бола алады. Жұмыста жүргізілген шолу мен салыстырмалы талдау негізінде деректерді интеллектуалды аналитикалық өңдеу процесін ұйымдастырудың қолданыстағы тәсілдері негізінде денсаулыққа зиян бойынша сақтандыру төлемдерінің мөлшерін одан әрі болжау үшін аналитикалық құралдардың критерийлері мен жіктеу жүйесі сыналды.

Кілт сөздер: *Data Mining,* кластерлеу, болжау, сызықтық регрессия, көпмүшелік регрессия, алгоритмдердің сапа көрсеткіші, машиналық ақыл, сақтандыру.

Кіріспе

Күрделі және серпінді ортада табысқа жету үшін компаниялар үнемі өзгеріп отыратын нарық жағдайларына бәсекелестеріне қарағанда жақсы бейімделуі керек, сондықтан стратегия ұғымы жақында танымал бола бастады. Қазіргі экономикада стратегия туралы өзіндік түсінік бар және ол кейбір аспектілерде тиісті анықтамалардан, сондай-ақ ұйымдасқан адам қызметінің басқа салаларында болатын жалпы қабылданған мөрлер мен стереотиптерден өзгеше болуы мүмкін (Michael J. et al., 1997).

Халықаралық тәжірибе көрсеткендей, өнім өндірушілердің қатаң бәсекелестігі жағдайында кәсіпкерлер мен менеджерлер тек өз тәжірибесіне, нұсқауларына немесе сәтті жағдайларына сене алмайды, өйткені кәсіпкерлік істің негізі кәсіпорынның стратегиясы мен дамуын жоспарлау болып табылады (Joshi M. et al., 2000). Тиімді стратегиялық басқарудың маңыздылығын түсінуге сүйене отырып, компанияның өнімділігін арттыру құралдарын іздеу менеджменттің заманауи міндеті екендігі айқын болады. Заманауи технологиялар клиенттердің, бәсекелестердің, контрагенттердің және т.б. тәртібіне аналитиканы сапалы жүргізуге мүмкіндік береді.

[*]Хат-хабарларға арналған автор. E-mail *denor1980@mail.ru.*

Компания қызметін басқарудың осындай жүйелерінің бірі *Data Mining* технологиясы болып табылады. Ол адам қызметінің көптеген салаларында көптеген мәселелерді шешеді. *Data Mining* банк ісі мен саудаға, медицина мен өндіріске, сақтандыру мен телекоммуникацияға және басқа салаларға біртіндеп енгізілуде.

Мысалы, бөлшек сауда кәсіпорындары бүгінде дүкен маркалы несие карталары мен компьютерленген бақылау жүйелерін қолдана отырып, әрбір жеке сатып алу туралы толық ақпарат жинайды. Бөлшек сауда саласында *Data Mining* көмегімен шешуге болатын типтік міндеттер:

– сатып алу себетін талдау (ұқсастықты талдау) тауарларды анықтауға арналған, себеттер жарнаманы жақсарту, тауарлар қорын құру стратегиясын және оларды сауда залдарында орналастыру тәсілдерін жасау үшін қажет.

– уақытша үлгілерді зерттеу сауда кәсіпорындарына тауарлық-материалдық құндылықтарды құру туралы шешім қабылдауға көмектеседі. Ол «Егер сатып алушы бүгін бейнекамера сатып алса, онда ол қанша уақыттан кейін жаңа батареялар мен пленканы сатып алуы мүмкін» сияқты сұрақтарға жауап береді (Kargupta, H. et al., 1999).

Банк ісі. *Data Mining* технологиясының жетістіктері банкте келесі жалпы міндеттерді шешу үшін қолданылады:

– несиелік карта бойынша алаяқтықты анықтау. Кейіннен алаяқтық болып шыққан бұрынғы транзакцияларды талдау арқылы банк мұндай алаяқтықтың кейбір стереотиптерін анықтайды.

– клиенттерді сегменттеу. Клиенттерді әртүрлі санаттарға бөлу арқылы банктер әртүрлі клиенттер топтарына әртүрлі қызмет түрлерін ұсына отырып, өздерінің маркетингтік саясатын мақсатты және тиімді етеді.

– клиенттердің өзгеруін болжау. *Data Mining* банктерге өз клиенттерінің болжамды құндылық модельдерін құруға және әр санатқа сәйкес қызмет көрсетуге көмектеседі.

Телекоммуникация. Телекоммуникация саласында *Data Mining* әдістері компанияларға бар тұтынушыларды ұстап тұру және жаңаларын тарту үшін маркетинг пен баға бағдарламаларын қарқынды түрде ілгерілетуге көмектеседі. Әдеттегі іс-шаралардың ішінде біз мыналарды атап өтеміз:

– қоңыраулардың егжей-тегжейлі сипаттамалары туралы жазбаларды талдау. Мұндай талдаудың мақсаты — өз қызметтерін пайдаланудың ұқсас стереотиптері бар клиенттердің санаттарын анықтау және бағалар мен қызметтердің тартымды жиынтығын әзірлеу;

– клиенттердің адалдығын анықтау. *Data Mining*-ті клиенттердің сипаттамаларын анықтау үшін пайдалануға болады, олар белгілі бір компанияның қызметтерін бір рет қолдана отырып, оған адал болып қалады. Нәтижесінде, маркетингке бөлінген қаражатты қайтарымы көп болатын жерде жұмсауға болады.

Text Mining мәтінді семантикалық талдауды, ақпараттық іздеуді және басқаруды жүзеге асырудың жаңа әдістерін қамтиды. Бұл тапсырманы жүзеге асыратын бағдарламалар қандай да бір жолмен табиғи адам тілімен жұмыс істеуі керек және талданатын мәтіннің семантикасын түсінуі керек. Кейбір *Text mining* жүйелер жолдағы ішкі жолды табуға негізделген.

Call Mining технологиясы сөйлеуді тануды, оны талдауды және *Data Mining*-ті біріктіреді. Оның мақсаты — операторлар мен клиенттер арасындағы келіссөздер жазбаларын қамтитын аудио мұрағаттарды іздеуді жеңілдету. Осы технологияның көмегімен операторлар клиенттерге қызмет көрсету жүйесіндегі кемшіліктерді анықтай алады, сатуды ұлғайту мүмкіндіктерін таба алады, сонымен қатар клиенттердің өтініштеріндегі тенденцияларды анықтай алады.

Веб-технологиялар. Web Content Mining «ақпараттық шу» шамадан тыс жүктелген әртүрлі интернет көздерінен сапалы ақпаратты автоматты түрде іздеуді және алуды білдіреді. Бұл сонымен қатар құжаттарды кластерлеу мен аннотациялаудың әртүрлі құралдары туралы.

Web Usage Mining веб-түйін пайдаланушысының немесе олардың тобының әрекеттеріндегі заңдылықтарды анықтауды білдіреді. Келесі ақпарат талданады:

– пайдаланушы қандай беттерді қарады;

– бетті қарау реті қандай.

Бизнестегі басқа қосымшалар. *Data Mining* көптеген басқа салаларда қолданылуы мүмкін:

– автомобиль өнеркәсібін дамыту. Автокөліктерді құрастыру кезінде өндірушілер әрбір жеке клиенттің талаптарын ескеруі керек, сондықтан оларға белгілі бір сипаттамалардың танымалдылығын болжау мүмкіндігі және әдетте қандай сипаттамаларға бірге тапсырыс берілетінін білу қажет;

– кепілдік саясаты. Өндірушілер кепілдік өтінімдерін беретін клиенттердің санын және өтінімдердің орташа құнын болжауы керек;

– жиі ұшатын клиенттерді ынталандыру. Авиакомпаниялар осы ынталандыру шараларымен көбірек ұшуға итермелейтін клиенттер тобын таба алады.

Осылайша *Data Mining* технологиялары бізге экономика мен қоғамдық қызметтің көптеген салаларында әртүрлі компаниялардың қызметін тиімдірек басқаруға мүмкіндік береді деп айта аламыз.

Сақтандыру қызметтерінің саласы да ерекше емес. Зерттеуіміздің мақсаты ретінде біз келесідей тұжырым жасауға болатын гипотезаны қоямыз: *Data Mining* технологиясы сақтандыру жағдайлары шеңберінде медициналық төлемдерді болжауға мүмкіндік береді. Бұл компанияға сақтандыру жағдайы орын алған жағдайда клиенттер үшін сақтандыру төлемдерін тиімдірек болжауға мүмкіндік береді. Біз *Data Mining* технологиясы сақтандыру компаниясына сақтандыру жағдайларын төлеуге болашақ ақшалай шығындарды жақсы болжауға мүмкіндік берді деген гипотезаны тексергіміз келеді.

Әдебиетке шолу

Data Mining тұжырымдамасының өзі машиналық оқыту, математикалық статистика және жасанды интеллект ұғымына негізделген. Сонымен қатар *Data Mining* бізге бизнесті басқару немесе әлеуметтік мәселелерді шешу саласында дайын шешімдер алуға мүмкіндік бермейтінін түсіну қажет. *Data Mining* технологиясының өзі деректерді алудың, талдаудың, оларды өңдеу әдістері мен әдістерінің тиімді жүйесін құруды білдіреді. Бұл өз кезегінде тиімді шешімдер қабылдауға мүмкіндік береді.

Көптеген зерттеушілер деректерді талдау үшін машиналық оқытуды қолдану мәселелерімен айналысты. Қазірдің өзінде зерттеушілер жасанды интеллект көмегімен деректерді өңдеудің тиімді әдістері мен алгоритмдерін іздеді. Мысалы Б. Уидроу (1985) өз жұмысында берілген шамалармен деректерді адаптивті өңдеуді жүзеге асыра алатын тиімді жүйелерді іздеді (Widrow B.S., Stearns D., 1985). Марвин Мински жасанды интеллекті ақылға қонымды кезеңде жаңа білім алудың және көптеген деректерді өңдеудің маңызды әдісі ретінде қарастырды. (Minsky M., 1974). Дж. Хопфилд (1985) өз еңбектерінде нейрондық желілерді жаңа білім құрудың құралы ретінде зерттеді. Нейрондық желі стохастикалық жүйелер туралы ақпарат алу механизміне айналды. Дж. Хопилд жұмыс нәтижелерінде биологиялық жүйелерді ақпаратты өңдеу элементтерінің байланысы ретінде қарастыруға мүмкіндік берді. (Hopfield J.J., Tank D.W., 1985). Д.А. Норман және Д.Э. Румелхарт (1975) өз еңбектерінде ақпаратты «когнитивті» өңдеу ұғымына сүйенді. Олар ақпараттың алушыға жетудің көптеген жолдары бар деп мәлімдеді. Ал алушы өз кезегінде оны өңдеудің және нәтиже алудың әртүрлі тәсілдеріне ие болады деген (Norman D., Rumelhart D., 1975). Ақпаратты жинау, өңдеу жүйесіне, сондай-ақ әдістер мен механизмдерге арналған әдебиеттерге осындай қысқаша шолудан көріп отырғанымыздай, XX ғасырдың екінші жартысында зерттеушілер нейротрансмиттер желілеріне, соның ішінде физиология саласына негізделген түпкілікті нәтижелерді алудың әмбебап жолдарын жасауға баса назар аударды.

Ақпараттық технологиялар дамыған сайын ақпараттың үлкен көлемін өңдеу мәселелеріне, яғни машиналық оқыту жүйесіне ауыса бастады. Дж.Р. Куинлен (Quinlan J.R., 1986) «Шешім ағашы» атты деректерді өңдеу әдістемесіне арналған еңбектерінде өте үлкен деректер жиынтығымен жұмыс істегенде, барлық мысалдарды қолдана отырып, шешім ағашын құру қиын болуы мүмкін деп негіздеді. Мүмкін болған жағдайда да, бұл деректерді пайдаланудың ең жақсы тәсілі болмауы. Сонымен қатар, бастапқы деректер жиынтығын үлгілерге бөлуге болады, «ағашты» әр ішкі жиынға салуға болады, содан кейін нақты оңтайландырылған «ағашты» алу үшін жеке «ағаштарды» ақылды түрде біріктіруге болады. Өзін-өзі ұйымдастыруға негізделген деректердің едәуір көлемін өңдеуге арналған жасанды нейрондық желілерге арналған жұмыстар пайда болады. Мұндай желілер деректерді өңдеу алгоритмін өздері жасайды, ақпаратты өңдеудің ең тиімді жолдарын іздейді (Kohonen T., 2001).

Data Mining теориясының негізін қалаушылардың бірі — Г. Пятицкий-Шапиро. 1989 жылы Г. Пятицкий-Шапиро семинарда деректерді өндіру саласын ұсынды. Ол *GTI Labs* компаниясының қызметкері болған кезде, ол белгілі бір ережелерді автоматты түрде табу мүмкіндігіне қызығушылық танытты, кейбір мәліметтер базасына сұраныстарды жеделдету үшін. Содан кейін екі термин енгізілді — *Data Mining* және *knowledge discovery In Data*. 1993 жылы «Knowledge Discovery Nuggets» алғашқы ақпараттық бюллетені енгізілді, бір жылдан кейін, 1994 жылы Григорий Пятецкий-Шапиро *Data Mining* бойынша алғашқы сайттардың бірін құрды (Piatetsky-Shapiro, G., 1991).

Әдістері

Сақтандыру жүйесіндегі *Data Mining* құралдары көптеген бағалау әдістеріне сүйене алады. Сонымен, *Data Mining* біздің зерттеуімізде сызықтық регрессия, көпмүшелік регрессия әдістері, алгоритмдердің деректер сапасының көрсеткіштері жатыр. Бұл зерттеу әдістерін қолдану көптеген авторлармен негізделген (Ian H., Witten et al, 2011; Han J., Kamber M., 2010).

Бұл әдістер бізге бірқатар факторларға байланысты сақтандыру жағдайларының басталу факторларын бағалауға мүмкіндік береді. Сызықтық регрессия мақсатты индикатордың (сақтандыру жағдайының басталуы) көптеген әсер ету факторларына тәуелділік дәрежесін анықтауға мүмкіндік береді. Сызықтық регрессияны қолданған кезде біз екі немесе одан да көп айнымалыларға тікелей тәуелділікте (түзу сызықты) сақтандыру жағдайының пайда болу факторларының әсер ету дәрежесін бағалай аламыз (Pawel Cichosz, 2015).

Сондай-ақ, зерттеу нәтижелерін растаудың балама нұсқасы үшін біз полиномдық регрессия негізінде сақтандыру жағдайларының басталу факторларының тәуелділік дәрежесін есептейміз (Hastie, T. et al, 2009; Liang Wang et al, 2009; Vapnik V.N., 1998).

Алынған нәтижелерге сүйене отырып, біз осы нақты жағдайда есептеу әдістерінің қайсысы тиімдірек екендігі туралы қорытынды жасай аламыз. Көпмүшелік регрессияның маңызды артықшылығы — бұл зерттеушілерге деректердің сызықтық емес тәуелділігі болса да, сенімді нәтижелерге қол жеткізуге мүмкіндік береді. Осылайша, сызықтық және көпмүшелік регрессияның дәйекті құрылысы бізге берілген тапсырмалар шеңберінде сызықтық емес деректерді зерттеудің жақсы нәтижесін беруге мүмкіндік жасайды.

Нәтижелер

Зерттеудің мақсаты — олардың өзара байланысын бақылау үшін әртүрлі сипаттамаларды зерттеу және болашақ медициналық шығындарды болжау үшін қолданылатын медициналық шығындармен салыстырғанда адамның жасы, физикалық/отбасылық жағдайы және орналасқан жері сияқты бірнеше сипаттамаларына негізделген бірнеше сызықтық регрессияны құру және олардың алынатын сыйлықақы туралы шешім қабылдауға әсері.

Деректер жиынының құрылымын сипаттау.

Үлгі 6 тәуелсіз сипаттамамен сипатталған 1338 нысаннан және сандық шкалада ұсынылған 1 мақсатты зарядтардан тұрады.

Төменде датасеттің құрылымы сипатталған:

– age – жасы.

– sex – жынысы (еркек, әйел).

– bmi-дене салмағының индексі, бойға қатысты салыстырмалы түрде жоғары немесе төмен, дене салмағының объективті индексі (кґ/м^2) бой мен салмақтың арақатынасын қолдана отырып, ең дұрысы 18,5-тен 24,9-ға дейін.

– children –медициналық сақтандыруы бар балалар саны;

– smoker – сақтанушы темекі шегеді ме?

– region –ҚР-да тұратын клиенттің аймағы;

– charges – медициналық шығындардың мөлшері.

Деректер жиынтығында айнымалылардың ешқайсысы бойынша бос орындар табылған жоқ, бірақ бір телнұсқа алынды.

Сипаттамалық статистика.

1-кестеде Dataset сандық белгілерінің сипаттамалық статистикасы туралы жиынтық ақпарат көрсетілген.

1-кесте. Dataset сипаттамалық статистикасы

	age	bmi	children	charges
mean	39.207025	30.663397	1.094918	1.592446e+06
std	14.049960	6.098187	1.205493	1.453201e+06
min	18.000000	15.960000	0.000000	1.346200e+05
25%	27.000000	26.296250	0.000000	5.688325e+05
50%	39.000000	30.400000	1.000000	1.125835e+06
75%	51.000000	34.693750	2.000000	1.996782e+06
max	64.000000	53.130000	5.000000	7.652450e+06
Ескерту – Авторлардың меншікті есептеулері негізінде құрылған				

Категориялық белгілер бойынша ақпарат 2-кестеде келтірілген.

2-кесте. Категориялық белгілердің таралуы

Белгілері	Бірегей тегтер саны	Кластары	Класс нысандарының саны	Нысандарда жиі кездесетін кластар
sex	2	еркек	676	еркек
		әйел	662	
smoker	2	жоқ	1064	жоқ
		бар	274	
region	4	солтүстік	364	солтүстік
		оңтүстік	325	
		батыс	325	
		шығыс	324	
Ескерту – Авторлардың меншікті есептеулері негізінде құрылған				

Сыныптың теңгерімсіздігі бар жалғыз белгі — smoker. Басқа белгілерде әртүрлі сыныптардың объектілері біркелкі ұсынылған.

Деректерді кодтау.

Деректерді талдау алгоритмдерінің басым көпшілігі сапалы белгілермен жұмыс істей алмайды, біз оларды сандарға айналдырамыз:

Екілік-екілік шкаладағы белгілердің көрінісі [0; 1]. Бинаризация sex («еркек»: 1, «әйел»: 0) және smoker («иә»: 1, «жоқ»: 0) белгілері үшін пайдаланылды.

Мақсатты кодтау — мақсатты атрибут арқылы категориялық айнымалыларды кодтау әдісі. Әрбір сынып сандық таргет жағдайында — медицинамен/орташамен, сапалық жағдайда — осы сыныпқа жататын үлестермен (ықтималдықпен) ауыстырылады. Бұл жағдайда аймақ белгісі әр аймақ бойынша шығындардың медианалық мәні арқылы кодталады (3-кесте).

3-кесте. Region атрибутын кодтау

Region белгісініңкластары	Медициналық шығындар медианасы
Шығыс	1206910
Батыс	1077230
Солтүстік	1115290
Оңтүстік	1055830
Ескерту – Авторлардың меншікті есептеулері негізінде құрылған	

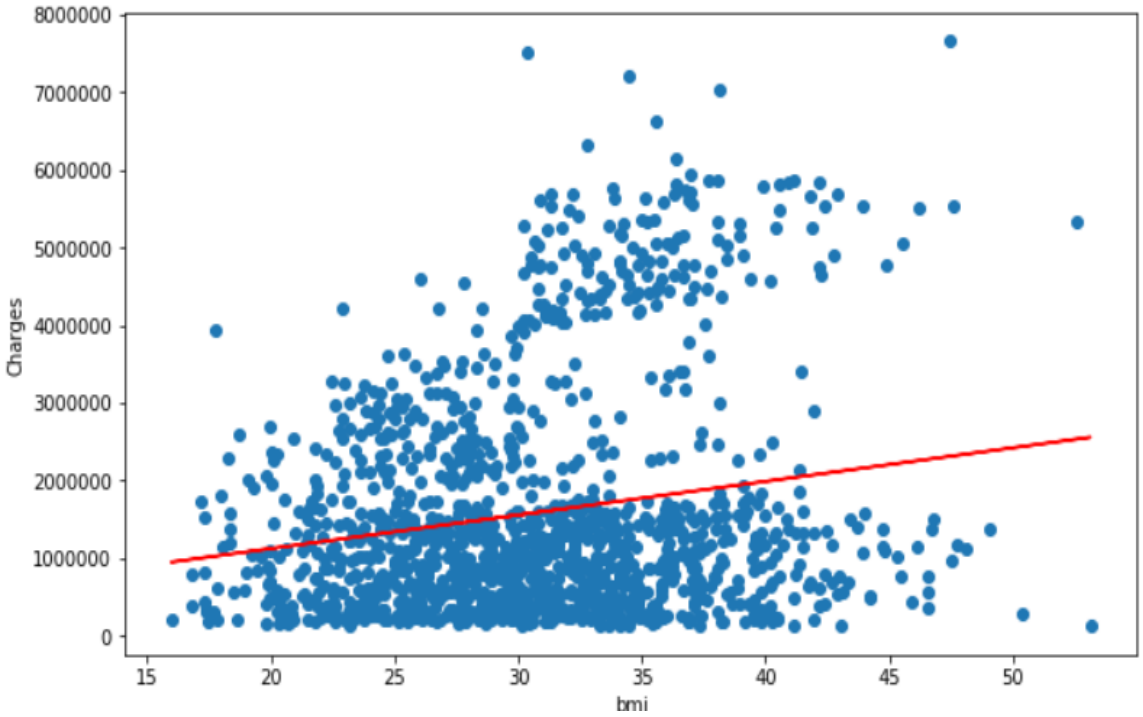
Факторлардың байланысын зерттеу.

Атрибуттардың өзара байланысын зерттеу үшін корреляциялық матрица құрылды (4-кесте). Корреляциялық матрица — M * M өлшемінің квадрат матрицасы, мұндағы M – негізгі диагональға қатысты симметриялы атрибуттар саны. Бұл жағдайда матрицада 7 жол және бірдей бағандар бар.

4-кесте. Корреляциялық матрица

	age	sex	bmi	children	smoker	median charges by region	charges
age	1.00	-0.02	0.11	0.04	-0.03	-0.00	0.30
sex	-0.02	1.00	0.05	0.02	0.08	0.00	0.06
bmi	0.11	0.05	1.00	0.01	0.00	-0.05	0.20
children	0.04	0.02	0.01	1.00	0.01	-0.03	0.07
smoker	-0.03	0.08	0.00	0.01	1.00	0.03	0.79
medianchargesbyregion	0.00	0.00	-0.05	-0.03	0.03	1.00	0.03
charges	0.30	0.06	0.20	0.07	0.79	0.03	1.00
Ескерту – Авторлардың меншікті есептеулері негізінде құрылған							

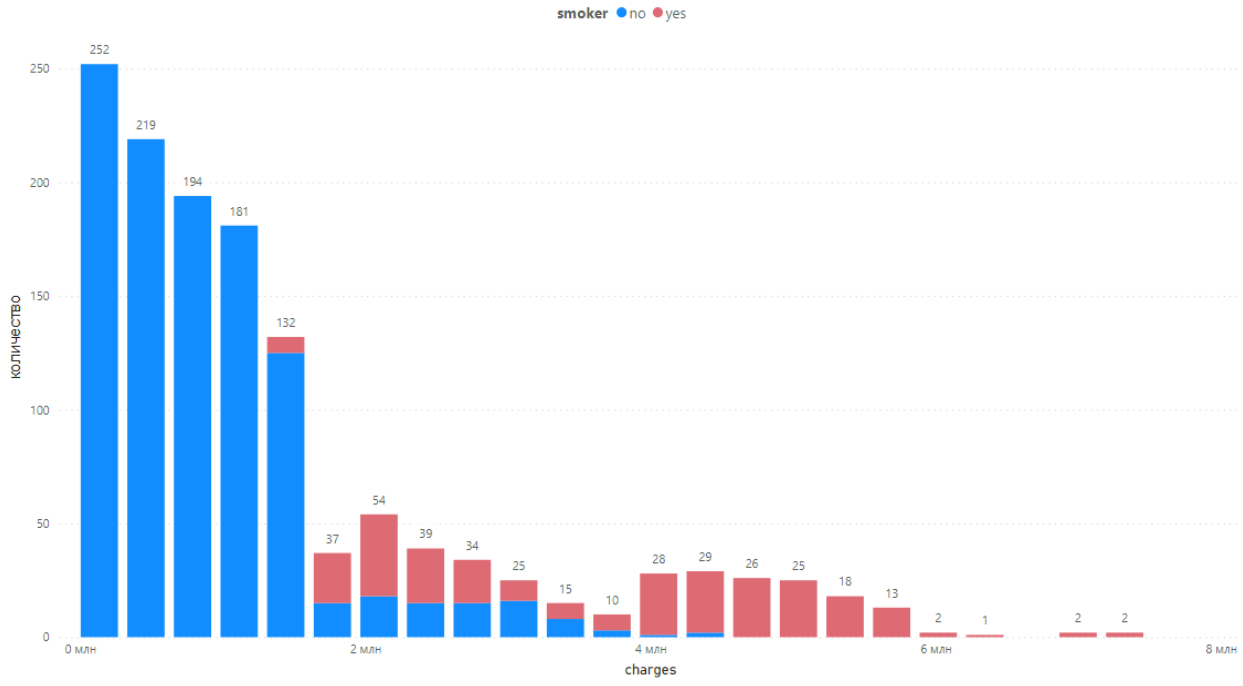
Барлық болжаушылар мақсатпен оң корреляцияға ие, бірақ тек smoker ғана күшті корреляцияға ие, әлсіз – age и bmi (1–сурет), өте әлсіз-жыныс белгілері, sex, children и median charges by region.



1-сурет. Bmi және charges белгілері арасындағы әлсіз корреляция

Ескерту – Авторлардың меншікті есептеулері негізінде құрылған

Сонымен қатар 2-суретте темекі шегетін клиенттердің медициналық шығындары әлдеқайда көп екенін көрсетіледі.



2-сурет. Smoker белгісі бойынша медициналық шығындар сомасын бөлу гистограммасы

Ескерту – Авторлардың меншікті есептеулері негізінде құрылған

Мақсат пен smoker белгісі (r = 0.7) арасындағы күшті корреляцияға және жоғарыда көрсетілген диаграмманың ерекшеліктеріне сүйене отырып, smoker шығындарды болжаудың негізгі факторы болып табылады.

Сызықтық көп регрессияның алғышарттарының бірі — тәуелсіз атрибуттар арасындағы коллинеарлықтың (сызықтық тәуелділіктің) болмауы. Алайда, эмпирикалық эконометрикалық зерттеулерде функционалды сызықтық тәуелділік жиі кездеспейді, бірақ екі немесе бірнеше тәуелсіз айнымалылар арасында жоғары корреляция болуы мүмкін. Бұл жағдайда олар мультиколлинеарлық проблема туралы айтады.

Тәуелсіз белгілер арасында күшті корреляция табылмады (мультиколлинеарлық проблеманың болмауы), сондықтан модельді оқытуға барлық болжаушыларды қосуға болады.

Алайда, егер сынақ деректерінде шығарындылар болса, онда модельдерді бір-бірімен объективті салыстыру қиын болады: шығарындылардағы қателер объектілердің негізгі жиынтығындағы қателіктердегі айырмашылықтарды жасырады.

Модельдердің сапасын құру және бағалау.

Сызықтық регрессия әдісіне негізделген модельдерді құру үшін көптеген мәліметтер оқыту (80%) және тестілік (20%) болып бөлінді. Барлығы 1-ден 11-ге дейінгі сызықтық және көпмүшелік регрессияның 11 моделі оқытылды.

Барлық модельдер sklearn сыныптары арқылы жасалған (linear_model.LinearRegression және sklearn.preprocessing.PolynomialFeatures).

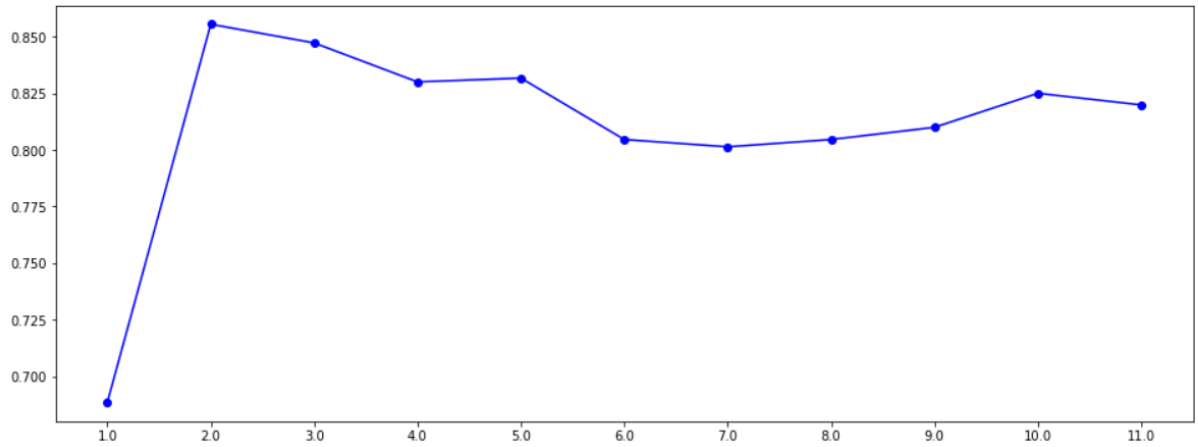
5-кестеде 10 кездейсоқ объектілердің 2-дәрежелі мәндерінің бірнеше сызықтық және көпмүшелік регрессиясының модельдері арқылы нақты және болжамды оннан онға дейін дөңгелектелген салыстырулар келтірілген.

5-кесте. Мақсатты айнымалының нақты және әртүрлі модельдер болжаған мәндерін салыстыру

Нысан номері	y_{true}	y_{pred} сызықтық регрессия көмегімен	y_{pred} 2-тәртіптегі полиномиалды регрессия көмегімен
1	1725920.0	1533369.0	1885023.0
2	4621390.0	3714415.0	4814917.0
3	542250.0	1012386.0	798458.9
4	727100.0	9017728.0	814058.6
5	569560.0	7184246.0	759926.9
6	2323430.0	3636001.0	2686599.0
7	1267250.0	1572586.0	1543243.0
8	302780.0	395202.9	446700.1
9	2765460.0	1402539.0	1376558.0
10	2013150.0	3378579.0	1721356.0
Ескерту – Авторлардың меншікті есептеулері негізінде құрылған			

Бұл жағдайда көп мүшелік модельдегі болжамдардың дәлдігі сызықтыққа қарағанда жоғары екенін байқау қиын емес.

Белгісіз жауап мәндерін жақсы болжайтын регрессиялық теңдеуді таңдайық (3-сурет). Модельдердің сапасын бағалаудың негізгі критерийі ретінде түсіндірілген R^2 дисперсиясының өлшемі қабылданды.



3-сурет. Регрессия теңдеуінің дәрежесіне байланысты R ^ 2 өзгеру графигі

Ескерту – Авторлардың меншікті есептеулері негізінде құрылған

Талқылау

Корреляциялық талдау негізінде бастапқыда мақсатты белгіні болжаудың негізгі факторы дұрыс атап өтілді. Бұл болжам бейнелеу әдісі арқылы да расталады. Бұл фактор регрессия теңдеуіндегі ең үлкен коэффициентке ие:

$$\hat{y} = -2102849 + 30072.36 * age + 13514.53 * sex + 36219.63 * bmi + 53395.91 * children + 2867290 * smoker + 68.21921 * median\ charges\ by\ region$$

Сызықтық регрессияның артықшылығы — оны оңай түсіндіруге болады. Бұл модельдің құрылымын бизнес пайдаланушыларға түсіндіру өте оңай.

Алайда, көпмүшелік регрессиялық модельдер мақсатты белгінің дисперсиясын әлдеқайда жақсы түсіндіреді. Сызылған графикке сәйкес екінші және одан да көп дәрежедегі R2 мәндері айтарлықтай ерекшеленбейтінін атап өткен жөн.

Болжау тұрғысынан жоғары ретті полиномдық регрессия негізінде оқытылған модельдер қайта оқытуға өте сезімтал. Бұл олардың үлгі үлгілерін өте жақсы сипаттайтынын білдіреді, бірақ популяция емес. Демек, іс жүзінде болжау үшін сызықтық регрессия және төмен ретті көпмүшелік регрессия модельдері қолданылады.

Қорытынды

Қазіргі уақытта өмірдің барлық салалары үлкен ақпарат ағынымен бетпе-бет келеді. Жинақталған деректер, олардың көлемі қаншалықты үлкен, адам оларды классикалық әдістермен талдай алмайды. Сонымен қатар, жаңа білімді іздеу адамның басты міндеті болып қала береді, өйткені бұл тек сапалы нәтижелерге қол жеткізуге мүмкіндік туғызады. Бүгінде жаңа білім іздеу жасанды интеллект пен ақпараттық технологияларға сүйене бастады. Ақпаратты өңдеу саласындағы тиімді құралдардың бірі *Data Mining* болды. *Data Mining* өзінің пәнаралық негізіне байланысты кеңінен қабылданды: оған жасанды интеллект, математикалық статистика, машиналық оқыту кіреді. *Data Mining*-тің маңызды ерекшелігі — ол дайын шешімдер бермейді. Бұл зерттеушілерге көптеген ақпаратты талдаудан сенімді деректер алуға мүмкіндік береді. *Data Mining*-тің маңызды ерекшелігі — ізделетін шаблондардың тривиалдығы. Бұл табылған үлгілер жасырын білім деп аталатын деректерді құрайтын деректердегі айқын емес, күтпеген (unexpected) заңдылықтарды көрсетуі керек дегенді білдіреді. Қазіргі қоғамдағы шикі деректер (raw data) білімнің терең қабатын қамтитыны туралы түсінік келді, оны сауатты талдау кезінде күтпеген нәтижелер анықталуы мүмкін, олардың есебі кез келген бизнестің, соның ішінде сақтандыру бизнесінің сәттілігінің кепілі болып табылады.

Сақтандыру саласы шешім қабылдау кезінде маңызды емес міндеттерге тап болады. Бұл сақтандыру жағдайларының басталуы көптеген факторларға байланысты болғандықтан орын алады. Олардың әсер ету дәрежесі үнемі өзгеріп отырады. Медициналық сақтандыру, сақтандыру қызметінің бір түрі ретінде, үнемі көптеген деректерді өңдеу қажеттілігіне тап болады. Жақында аналитиктер жақсы нәтиже алу үшін регрессияны талдауды және оның құралдарын қолдана бастады. Бұл әдіс осы көрсеткіштің басқа көрсеткіштерге тәуелділік дәрежесін алуға мүмкіндік береді, олардың саны мен сапасы әр түрлі болуы мүмкін.

Талдау үшін біз денсаулықты сақтандыру жағдайларының басталуының бірқатар факторларын алдық. Болашақ медициналық шығындарды анықтау үшін сақтандырылған адамның денсаулық деңгейіне әсер ететін бірқатар факторларды атап өту қажет болды: жас, темекі шегу, дене салмағының индексі, жынысы, медициналық сақтандыруы бар балалардың болуы, тұрғылықты жері және жұмсалған медициналық шығындардың мөлшері. Сызықтық регрессия функциясын қолдана отырып, біз сақтандыру жағдайының басталуының жоғарыда аталған факторларға тәуелділік дәрежесін анықтауға тырыстық. Алынған нәтижелер де тексерілді көпмүшелік регрессиялық модельдер.

Біздің талдауымыз көрсеткендей, жас пен темекі шегу денсаулыққа қатысты сақтандыру жағдайының деңгейіне үлкен мән береді. Бұл сақтандыру компаниясының клиенттерінің әл-ауқатына ең үлкен теріс әсер етуі мүмкін. Дене салмағының индексі де үлкен мәнге ие. Тұру орны және жасы да денсаулық сапасына әсер етеді. Мұның бәрі сақтандыру компаниясының басшылығына сақтандыру өтемақыларын төлеу шығындарын нақты жоспарлауға мүмкіндік жасайды. Сондай-ақ, сақтандыру жағдайының ықтималдығына әсер ету дәрежесін анықтау сақтандыру компаниясына клиентке келтірілген залалды өтеу тетіктері сипатталатын шарттарды неғұрлым икемді жасауға

мүмкіндік береді. Мысалы, егер адам темекі шегетін болса және оның денсаулығының нашарлау ықтималдығы жоғары болса, сақтандыру компаниясы аз төлемдер тағайындай алады немесе сақтандыру шартының құнын арттыра алады. Немесе денсаулықтың нашарлау факторын «жыныс» деп есептемеуге болады, бұл бұрынғы сақтандыру төлемдерінің мөлшеріне айтарлықтай және сақтандыру компаниясының болашақ шығындарына үлкен әсер етпейді.

Data Mining және бизнес-процестерді математикалық қамтамасыз ету компания басшылығы тарапынан болжаудың, басқарудың және бақылаудың ажырамас бөлігіне айналады. *Data Mining*-тің кең құралдары компаниялардың шығындарын тиімдірек басқаруға мүмкіндік берді, кластерлеу бизнес-процестерді жеке элементтерге бөлуге және олардың тиімділігін есептеуге мүмкіндік жасайды.

Әдебиеттер тізімі

Cichosz, P. (2015). Data Mining Algorithms: Explained Using. R. Wiley.

Han, J., & Kamber, M. (2010). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 2nd ed. – Springer-Verlag.

Hopfield, J.J., & Tank, D.W. (1985). "Neural" computation of decisions in optimization problems. *Biol. Cybern*., *52*, 141–152.

Ian, H. Witten, Eibe Frank, Mark, & A. Hall (2011). Data Mining: Practical Machine Learning Tools and Techniques. "Morgan Kaufmann Publishers Inc.", P. 664.

Joshi, M., Karypis G., & Kumar, V. (2000). ScalParC: A New Scalable and Efficient Parallel Classification Algorithm for Mining Large Datasets. *Research Gate*, *3*, 121-130.

Kargupta, H., Hamzaoglu, I., & Stafford, B. (1999). Distributed data mining using an agent based architecture. In Proceedings of Knowledge Discovery and Data Mining.AAAI Press.

Kohonen, T. (2001). "Self-Organizing Maps". 3rd Edition. Springer-Verlag, Berlin, Heidelberg, New York.

Liang Wang, Li Cheng & Guoying Zhao (2009). Machine Learning for Human Motion Analysis. IGI Global.

Michael, J. Berry, A. & Gordon Linoff (1997). Data Mining Techniques. Jhon Wiley & Sons, Inc.

Minsky, M. (1974). A framework for representing knowledge. Cambridge.

Norman, D., & Rumelhart, D. (1975). Explorations in Cognition. San Francisco: Freeman Norman, D.A., Rumelhart, D.E., & the LNR Research Group.

Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89. *Workshop AI Magazine, 11(5),* 68–70.

Quinlan, J.R. (1986). Induction of Decision Trees. *Machine Learning. Kluwer Academic Publishers, 1,* 81–106

Vapnik, V.N. (1998). Statistical learning theory. N.Y.: John Wiley & Sons, Inc.

Widrow, B., & Stearns, S. D. (1985). Adaptive Signal Processing. Prentice-Hall, Englewood Cliffs.

Н.Н. Гелашвили, А. Сабыржан, Б.Х. Раимбеков, Г.А. Кенешева, Г.К. Абдраманова

Практическое использование методов интеллектуального анализа данных в системе медицинского страхования

Аннотация

Информация стала важнейшим фактором успешности деловой деятельности компании. От ее наличия, качества и своевременности зависит уровень конкурентоспособности фирмы. При этом сегодня количество поступающей информации огромное и возникает необходимость ее эффективной обработки.

Целью исследования является применение математических способов обработки информации с использование инструментария *Data Mining* для прогнозирования будущих страховых выплат по здоровью.

Методы: Для прогнозирования размера будущих страховых выплат и для выявления факторов наступления страховых случаев нами были выбраны методы линейной и полиномиальной регрессии и метрика алгоритмов. В качестве ключевых факторов были использованы пол и возраст клиентов, наличие привычки курить, соотношение массы тела к росту, место проживания, наличие детей с уже имеющейся страховкой, размер медицинских расходов и место проживания.

Результаты: На уровень наступления страхового случая по здоровью наибольшее значение имеют возраст и курение. Именно это может оказывать наибольшее негативное влияние на самочувствие клиентов страховой компании. Также достаточно большое значение имеет индекс массы тела. Менее всего на качество здоровья влияют место проживания и возраст. Все это позволяет руководству страховой компании более четко планировать собственные расходы на выплаты страховых компенсаций. Также выявление степени влияния на вероятность наступления страхового случая позволяет страховой компании более гибко составлять договоры, где будут описаны механизмы возмещения представленного ущерба клиенту.

Выводы: *Data Mining* может стать эффективным инструментом для эффективного управления деятельностью компании. На основе проведенного в работе обзора и сравнительного анализа инструментальных средств и существующих подходов к организации процесса интеллектуальной аналитической обработки данных была апробирована система критериев и классификации аналитических инструментов для дальнейшего прогнозирования размера страховых выплат по ущербу по здоровью.

Ключевые слова: *Data Mining*, кластеризация, прогнозирование, линейная регрессия, полиномиальная регрессия, метрика качества алгоритмов, машинный разум, страхование.

Gelashvili N.N., Sabyrzhan A., Raimbekov B.Kh., Kenesheva G.A., Abdramanova G.K.

Practical use of data mining techniques in the health insurance system

Abstract:

Information has become the most important factor in the success of the company's business. The level of competitiveness of the company depends on its availability, quality and timeliness. At the same time, today the amount of incoming information is huge and the need arises for its effective processing.

Object: to apply mathematical ways of processing information using the Data Mining toolkit to predict future health insurance benefits.

Methods: to predict the size of future insurance payments and to identify the factors of the occurrence of insurance claims, we chose the methods of linear and polynomial regression of the algorithm metric. The key factors used were: gender and age of customers, smoking habit, weight-to-height ratio, place of residence, presence of children with existing insurance, size of medical expenses and place of residence.

Findings: Age and smoking are most important for the level of occurrence of an insured health event. This is what can have the greatest negative impact on the well-being of customers of the insurance company. Body mass index is also quite important. Least of all, the quality of health is influenced by the place of residence, and age. All this allows the management of the insurance company to more clearly plan their own expenses for the payment of insurance compensation. Also, the identification of the degree of influence on the likelihood of an insured event allows the insurance company to more flexibly draw up contracts, which will describe the mechanisms for compensation for the presented damage to the client.

Conclusions: Data Mining can become an effective tool for effective management of the company's activities. Based on the review and comparative analysis of tools and existing approaches to the organization of the intelligent analytical data processing process, a system of criteria and classification of analytical tools was tested to further predict the amount of health damage insurance benefits.

Keywords: Data Mining, clustering, forecasting, linear regression, polynomial regression, algorithm quality metric, machine mind, insurance.