Decision Trees as a Business Online Advertising Strategy Optimization Tool

Vesela Mihova, PhD CandidateD
Faculty 'Natural Sciences and Education'
Department of Applied Mathematics & Statistics
University of Ruse 'Angel Kanchev', Bulgaria
E-mail: vmicheva@uni-ruse.bg

Дърво на решенията като инструмент за оптимизация на стратегията за онлайн бизнес реклама

Докторант Весела Михова
Факултет „Природни науки и образование"
Катедра „Приложна математика и статистика"
Русенски университет „Ангел Кънчев", България
E-mail: vmicheva@uni-ruse.bg

Abstract: *Online advertising services such as Google AdWords provide their customers with statistics on the performance of their ads. Based on this data, each company could optimize its ads. In the work presented, the potentials of using decision trees for classification in the field of online advertising are outlined by solving a problem from practice. A subsequent analysis shows which actions the company could take to optimize its online advertising strategy. The algorithm proposed could be used for other data in a similar situation.*

Key words: *Google AdWords, Online Advertising Strategy, Decision Tree.*

Резюме: *Онлайн рекламните услуги като Google AdWords предоставят на своите клиенти статистически данни за ефективността на техните реклами. Въз основа на тези данни всяка компания може да оптимизира рекламите си. В настоящата работа, чрез решаване на проблем от практиката, са очертани възможностите за използване на дърво на решенията с цел класификация в областта на онлайн рекламата, както и действията, които да се предприемат, за оптимизиране на онлайн рекламната стратегия.*

Ключови думи: *Google AdWords, стратегия за онлайн реклама, дърво на решенията.*

JEL Classification: *M21, M37*

I. Introduction

Nowadays, in order to achieve good realization on the market, a product or a service should not only be affordable and of good quality, but also have an effective advertising strategy. Competitive and dynamic markets require private companies to have an adequate advertising strategy. For this purpose, business often resorts to internet services, as it could reach through the network a huge number of potential buyers within a short time. In the most advanced economies, Internet advertis-

I. Въведение

В днешно време, за да има един продукт или услуга добра реализация на пазара, той трябва не само да е качествен и на добра цена, но и да има ефективна рекламна стратегия. Динамиката на пазара на стоки и услуги и засилената конкуренция в условията на пазарна икономика изискват частните фирми да имат адекватна рекламна стратегия. За тази цел бизнесът често прибягва до интернет услуги, тъй като в мрежата може за кратко време да достигне до огромно количество потенциални купувачи. В найing costs reach 50% of the total advertising costs of the companies (Abcbg.com, 2017). Among the main advantages of advertising on the Internet are good accountability (statistics about interest, attitudes and behavior of the consumers) and efficiency (lower cost and greater flexibility).

The business online advertising strategy typically includes Google ads (in Google Search Engine), website banners, social media ads, presence in the electronic media, development of own business blog, internet videos, electronic word-of-mouth, etc. The combination of different types of advertising and the allocation of resources (time, money) to them depends on what budget the company has allocated to advertising as a whole, the target market of the company's products or services, and the specifics of these products and services.

The research object of the current work are the ads that appear when searching for keywords in Google's search engine. The subject of the research is the effectiveness of these ads, and the goal is to illustrate the use of decision tree as a tool that companies can apply to optimize their online advertising strategy. To accomplish the stated goal, the following tasks have been solved:

- The decision tree method has been used in order to classify company's online ads according to their conversion rate;
- Two different approaches of validating the decision tree have been considered: split-sample validation and cross-validation, and the results of both methods have been compared;
- An analysis has been carried out – which actions the company could take to optimize its online advertising strategy.

The data for the empirical research is provided using the online advertising services of Google AdWords. With Google AdWords, businesses can reach relevant customers on their preferred websites anywhere on

развитите икономики разходите за интернет реклама достигат 50% от общите рекламни разходи на компаниите (Abcbg.com, 2017). Сред основните предимства на рекламата в Интернет са добрата отчетност (статистики относно интереса, нагласите и поведението на потребителите) и ефективността (по-ниска цена и по-голяма гъвкавост).

Онлайн рекламната стратегия на фирмите обикновено включва реклами в Google (показват се при търсене на ключови думи в търсачката), банери в уебсайтове, реклами в социалните медии, присъствие в електронните медии, разработка на собствена медия под формата на блог, интернет клипове, електронно предаване „от уста на уста" и др. Комбинацията от различните видове реклама и разпределението на ресурси (време, пари) към тях зависят от това какъв бюджет е заделила компанията за реклама като цяло, от целевия пазар, на който фирмата иска да реализира продуктите или услугите си, както и от спецификата на самите продукти и услуги.

В настоящата работа обект на изследване са рекламите, които се показват при търсене на ключови думи в търсачката на Google. Предмет на изследването е ефективността на тези реклами, а целта е да се онагледи използването на дърво на решенията като инструмент, който компаниите могат да прилагат за оптимизация на онлайн рекламната си стратегия. За реализиране на поставената цел са решени следните задачи:

- с помощта на метода дърво на решенията е направена класификация на онлайн рекламите на една компания според нивото им на реализация;
- приложени са два различни подхода за валидиране на дървото на решенията: валидиране с разделен подбор и крос-валидация, като резултатите от двата метода са сравнени;
- извършен е анализ на това какви действия може да предприеме фирмата, за да оптимизира онлайн рекламната си стратегия.

Нужните данни за емпиричното изследване са набавени с помощта на услугата за онлайн рекламиране Google AdWords. Благодарение на Google AdWords фирмите могат да достигнат до подходящи клиенти на предпочитани от тях уебсайтове the web (Adwords.google.com, 2017), or attain Google Search Engine users (the ad is shown when searching for keywords).

Google AdWords shows how many people have noticed an ad and what percentage of them have clicked to visit the website the add refers to or have called the ad phone number. The platform also allows tracking the actual sales generated by the advertiser's website as a direct result of its ads.

Hereinafter, the following definitions have been used (Support.google.com, 2017)

- *Click* - when someone clicks an ad.
- *Impression* - each time an ad is shown on a search result page or other site on the Google Network is counted.
- *Clickthrough Rate* - the number of clicks that an ad receives divided by the number of times it is shown (Clicks / Impressions = CTR), measured in percentages.
- *Average Position* - describes how an ad typically ranks against other ads. The highest position is "1", and there is no "bottom" position. An average position of 1-8 is generally on the first page of search results, 9-16 is generally on the second page, and so on. Average positions can be between two whole numbers. For example, an average position of "1,6" means that the ad usually appears in positions 1 or 2.
- *Conversion* – it happens when someone clicks an ad and then takes an action, which the advertising company has defined as valuable for its business, such as an online purchase or a call to the business from a mobile phone.
- *Conversion Rate* - the average number of conversions per ad click, shown as a percentage.

Google AdWords provides its customers with statistics on the performance of their ads. Based on this data, each company could optimize its ads, test new search keywords, stop advertising temporarily, and so on.

Mathematical tools that are part of the so-

навсякъде в мрежата (Adwords.google.com, 2017). Или до потребителите на търсачката на Google (рекламата се показва при търсене на ключови думи).

Google AdWords показва колко хора са забелязали дадена реклама и какъв процент са кликнали, за да посетят уебсайта, към който тя препраща, или са се обадили на телефонния номер от рекламата. Платформата позволява и проследяване на действителните продажби, които генерира уебсайтът на фирмата-рекламодател като пряко следствие от рекламите ѝ.

По-нататък в настоящата работа са използвани следните дефиниции (Support.google.com, 2017):

- *Кликване* - когато някой щракне върху рекламата;
- *Импресия* - отчита се всеки път, когато дадена реклама се показва на страница с резултати от търсенето или на друг сайт в мрежата на Google;
- *Честота на кликване* - броят на кликванията, които дадена реклама получава, разделен на броя показвания на рекламата (кликвания / импресии = честота на кликване). Измерва се в проценти.
- *Средна позиция* – описва как дадена реклама обикновено се нарежда спрямо други реклами. Най-високата позиция е „1" и няма най-ниска позиция.
- *Брой реализации* – реализация се осъществява, когато потребител кликне върху дадена реклама и след това предприеме действие, което рекламодателят е определил като ценно за бизнеса си като например онлайн покупка или обаждане до офиса на фирмата от мобилен телефон.
- *Ниво на реализация* - Средният брой реализации за кликване върху реклама, показан като процент.

Google AdWords предоставя на клиентите си статистики за представянето на техните реклами. На база на тези данни всяка фирма може да оптимизира рекламите си, да изпробва нови ключови думи за търсене, да спре рекламата си временно и т.н.

За оптимизиране на онлайн рекламни стратегии на помощ идват математически called "Data Mining" process are often used in order to optimize an online advertising strategy. These tools include descriptive analysis, link analysis, multi-dimensional statistical analysis, decision trees, forecasting, neural networks, and more.

Data Mining is used to extract useful information from large datasets and to display it in easy-to-interpret visualizations (Song and Ying, 2015). It is an interdisciplinary area that arises and develops on the basis of "neighboring" fields such as Applied Statistics, Machine Learning, Artificial Intelligence, etc. (Ivanov, 2016). Data Mining is more pragmatic than theoretical (Zaki, Meira and Meira Jr, 2014), (Olson and Delen, 2008). Its technology is based on the concept of development of templates reflecting multi-dimensional relationships in data (Han and Kamber, 2006). Among the main thematic issues that could be solved with Data Mining tools are the tasks of classification, clustering, forecasting and prediction, association, visualization, identification and analysis of deviations, assessment, link analysis, aggregation (Ivanov, 2016). Data Mining is applied in a number of areas, including marketing, sales, customer relationship management and behavioral models, for which extensive information is provided by Linoff and Berry (2011), Ngai and Chau (2009), Shaw, Subramaniam, Tan and Welge (2001) and others.

Decision trees, which were introduced in the 1960s, are one of the most effective methods for Data Mining. They represent a nonparametric method for analyzing the target variable as a function of explanatory characteristics (Filipov, 2014). They use the following algorithm: a dichotomous tree is constructed from nodes; at each node the population is subdivided into sub-populations based on the function of one of the variables; the system takes into account all possible divisions and chooses the best (resulting in minimal error of discrimination of the possible values of the target variable); the process continues until in each node has left records with only one of the possible values of the target variable, or until further separation

инструменти, които са част от така наречения „интелигентен анализ на данни": дескриптивен анализ, анализ на връзките, многомерен статистически анализ, дървета на решенията, прогнозиране, невронни мрежи и други. Интелигентният анализ на данни се използва за извличане на полезна информация от големи масиви от данни и за онагледяването ѝ в лесни за интерпретиране визуализации (Song & Ying, 2015). Той представлява интердисциплинарна област, възникнала и развиваща се на базата на „съседни" области като приложна статистика, машинно обучение, изкуствен интелект и др. (Иванов, 2016). Интелигентният анализ на данни има по-скоро прагматична, отколкото теоретична насоченост (Zaki, Meira & Meira Jr, 2014), (Olson & Delen, 2008). Технологията на анализа се основава на концепцията за разкриването на шаблони, отразяващи многоаспектни отношения в данните (Han & Kamber, 2006). Сред основните тематични проблеми, които могат да бъдат решени с инструменти на интелигентния анализ на данни, са задачите за класификация, клъстеризация, прогнозиране и предвиждане, асоциация, визуализация, идентификация и анализ на отклоненията, оценяване, анализ на връзките, обобщаване (Иванов, 2016). Интелигентният анализ на данни намира приложение в редица области, сред които маркетинг, продажби, управление на взаимоотношенията с клиенти и поведенчески модели, за което обширна информация дават Linoff & Berry (2011), Ngai & Chau (2009), Shaw, Subramaniam, Tan & Welge (2001) и други.

Въведените през 60-те години „дървета на решенията" са един от най-ефективните методи за интелигентен анализ на данни. Те представляват непараметричен метод за анализ на целевата променлива като функция от обясняващи характеристики (Филипов, 2014). Работят по следния алгоритъм: изгражда се дихотомно дърво от възли, като при всеки възел популацията се разделя на подпопулации на базата на функция на една от променливите; системата взима предвид всички възможни деления и избира най-доброто (което се получава при минимална грешка на дискриминиране на възможните стойности на целевата променлива); процесът продължава, докато във всеки възел не останат записи само is impossible (Filipov, 2014).

Decision trees are widely used in a number of disciplines (Hastie, Tibshirani and Friedman, 2009), because they are easy to work with, give unambiguous results, and are persistent even in the presence of missing data. Similar information, concerning the application of decision trees in personalized advertisements on internet storefronts, is provided by Kim, Lee, Shaw, Chang and Nelson (2001).

The decision tree method is used in the current work in order to classify a company's online ads according to their conversion rate. The classification has been done using SPSS (Science and Analytics, 2017). A subsequent analysis of the actions the company could take to optimize its online advertising strategy has been carried out.

Basic steps for statistical analysis with SPSS are presented from Goev (1996), Manov (2001), Pavlov and Mihova (2016). More specific information about the decision tree development process in SPSS is given by Magidson (2005), Baizyldayeva, Uskenbayeva and Amanzholova (2013) and others.

II. Exposition

AA database from an insurance company has been used for the purposes of the study. It contains information about the online advertising of the products of the company for the period March 2008 - January 2017 (monthly data). The company offers the following insurance products

- Property Insurance: Real Estate, CASCO, CARGO, Insurance of Construction and Assembly Works
- Liability Insurance: General Liability Insurance, Compulsory Civil Liability Insurance for Dangerous Activities, Professional Liability Insurance
- Legal Expenses Insurance
- Financial Risk Insurance
- Travel Assistance Insurance.

The available data includes 339 entries, with some missing monthly data for some of the types of insurance.

с една от възможните стойности на целевата променлива или докато по-нататъшно разделяне не е възможно (Филипов, 2014).

Дърветата на решенията намират широко приложение в редица дисциплини (Hastie, Tibshirani & Friedman, 2009), защото те са лесни за използване, дават недвусмислени резултати и са устойчиви дори при наличие на липсващи данни. Информация, близка до разглежданата тема, относно приложението на дървета на решенията в персонализирани онлайн реклами, дават Kim, Lee, Shaw, Chang & Nelson (2001).

В настоящата работа с цел класификация на онлайн рекламите на една компания според нивото им на реализация е използван методът дърво на решенията. Класификацията е направена с помощта на SPSS (Science and Analytics, 2017), след което е извършен анализ на това какви действия може да предприеме фирмата, за да оптимизира онлайн рекламната си стратегия. Основни стъпки за статистически анализ със SPSS са представени от Гоев (1996), Манов (2001), Павлов & Михова (2016). По-конкретна информация относно процедурата по разработка на дърво на решенията в SPSS дават Magidson (2005), Baizyldayeva, Uskenbayeva & Amanzholova (2013) и други.

II. Изложение

За нуждите на изследването е използвана база данни от застрахователна компания, съдържаща информация за онлайн реклама на застрахователните продукти на фирмата за периода март 2008 – януари 2017 г. (месечни данни). Компанията предлага следните застрахователни продукти:

- Имуществени застраховки: на недвижими имоти, КАСКО, КАРГО, на строително-монтажни работи (СМР);
- Застраховки на отговорности: обща гражданска отговорност (ГО), задължителна ГО за опасни дейности, професионални отговорности;
- Застраховане на правни разноски;
- Застраховки срещу финансов риск;

A classification of the types of ads (according to the type of insurance) has been done, based on the conversion rate as the main criteria and on the average position as an additional criterion. The idea of including the average position in the decision tree is that it has the following influence: the cases with higher average positions have a greater impact on the classification, those with lower average positions have less impact. The logic is that the cases with higher positions are shown backward in search and if these cases (or ads) have a high conversion rate, then the ad is really good - it leads to a high conversion of the product (although it is shown backward in search) and costs less to the company.

The Chi-squared Automatic Interaction Detection (CHAID) technique (Kass, 1980) has been used for the classification. An important note here is that unlike regression analysis, the CHAID technique does not require the data to be normally distributed.

In order to check the stability of the classification (whether it could be applied to the entire set), it is good practice to test how it works on a sample that was not used in the development of this classification. That is, to validate the results. Two different approaches of validating the decision tree have been considered in this work: split-sample validation and cross-validation. The results of the two methods have been compared, and then it has been analyzed which types of insurance have a high conversion rate and for which types change in the advertising strategy is required due to their low conversion rate.

Split-Sample Validation

With split-sample validation, the model is generated on a random sample of the data (development sample) and tested on the rest of the data (validation sample).

The development (training) sample normally includes 80% randomly selected records from the available database and the validation (test) sample includes the remaining 20% (in some cases, a variable is chosen to split the observations in the two samples). For the available data, such

- Помощ при пътуване (асистанс).

Наличната информация включва 339 записа, като има месеци, за които липсват данни за някои от видовете застраховки. Направена е класификация на видовете реклами (според типа на застраховката), базираща се на нивото на реализация като основен критерий и на средната позиция като допълнителен критерий. Идеята на включването на средната позиция в дървото на решенията е следната: случаите с по-висока средна позиция имат по-голямо влияние върху класификацията, тези с по-ниска – по-малко. Логиката е, че случаите с по-висока позиция се показват по-назад в търсенето и ако те имат високо ниво на реализация, значи рекламата наистина е добра (води до висока реализация на продукта, въпреки че излиза по-назад в търсенето и съответно от фирмата са имали по-малко разходи за нея).

За класификацията е използвана техниката CHAID (Kass, 1980). За разлика от регресионния анализ, тази техника не изисква данните да са нормално разпределени.

С цел да се провери дали една класификация е стабилна (приложима върху цялото множество), е добре да се изпробва как тя работи върху извадка, която не е използвана при разработката ѝ. Тоест, да се направи валидиране на резултатите. В настоящата работа са приложени два различни подхода за валидиране на дървото на решенията: валидиране с разделен подбор и крос-валидация. Резултатите от двата метода са сравнени, след което е направен анализ кои видове застраховки имат високо ниво на реализация и за кои е нужна промяна в рекламната стратегия поради ниската им реализация.

Валидиране с разделен подбор

При валидирането с разделен подбор моделът се генерира върху случайна извадка от данните, наречена "работна" и се изпробва върху останалата част от данните, които влизат в така наречената "тестова" извадка.

Обикновено разпределението на наблюденията е 80%:20% за работната срещу тестовата извадка (или пък се избира променлива, по която да се разделят на-a ratio (80%:20%) results in a very small number of observations in the test sample (68 cases), which in turn may lead to a distortion of the results. For this reason, the development sample in the presented work includes 75% randomly selected cases. This sample has been used to classify the data by the selected variables. The validation sample includes the remaining 25% - it has been left to validate the results.

As a result of the classification, generated on the development sample, the observations have been divided into two groups according to the conversion rate of the advertised products (Fig. 1)

блюденията). Тук такова съотношение води до много малък брой наблюдения в тестовата извадка (общо 68), което от своя страна може да доведе до изкривяване на резултатите. Поради тази причина за класифицирането на данните по избраните признаци е използвана извадка от 75% от записите, подбрани на случаен принцип. Останалите 25% са оставени за валидиране на резултатите.

В резултат от класификацията, генерирана върху работната извадка, наблюденията са разделени в две групи според нивото на реализация на отделните рекламирани продукти (Фиг. 1):

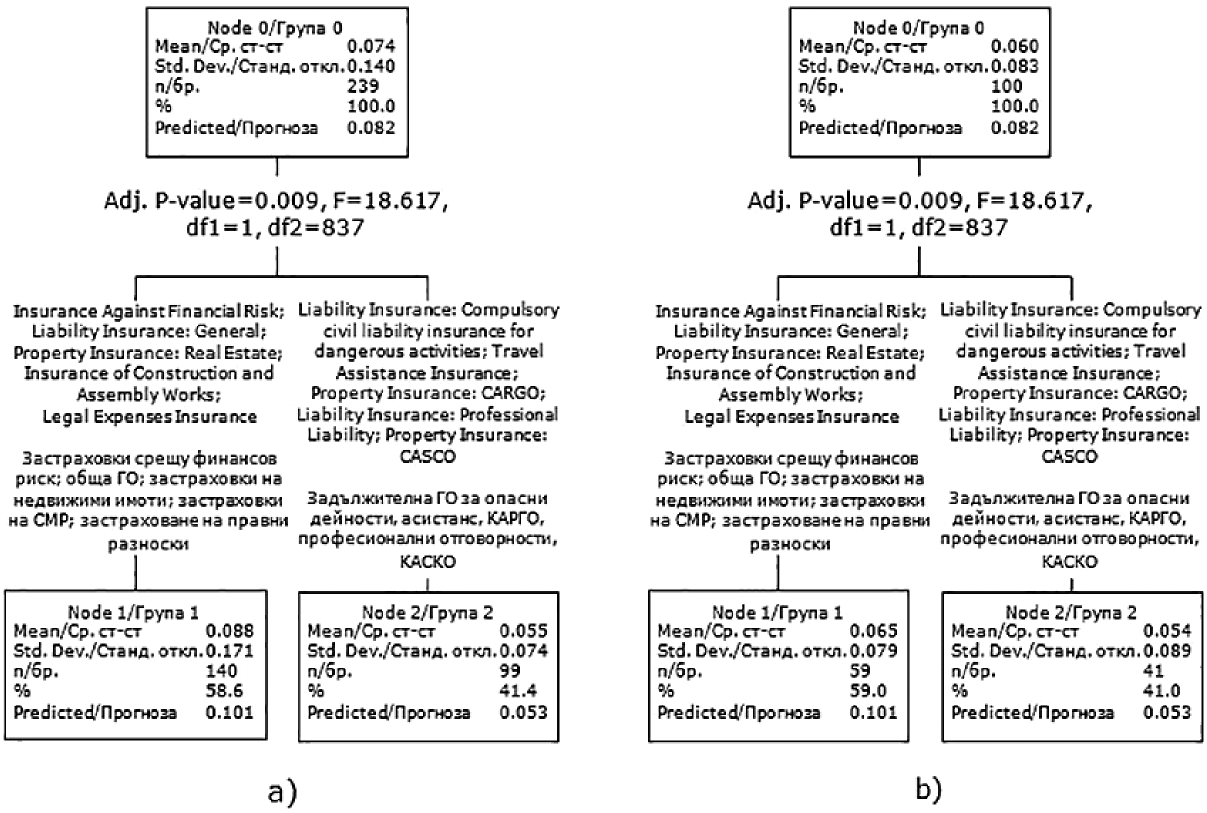

Figure 1. Decision tree: (a) development sample and (b) validation sample

Фигура 1. Дърво на решенията: (а) работна извадка и (б) валидационна извадка

Source: *Own Calculations with SPSS /* ***Източник:*** *Собствени изчисления със SPSS*

- Group 1 (Node 1): Financial Risk Insurance, General Liability Insurance, Real Estate Insurance, Insurance of Construction and Assembly Works, Legal Expenses Insurance
- Group 2 (Node 2): Compulsory Civil Liability Insurance for Dangerous Activities, Travel Assistance Insurance, CARGO, Professional Liability Insurance, CASCO.

- Група 1: застраховки срещу финансов риск, обща ГО, застраховки на недвижими имоти; застраховки на СМР, застраховане на правни разноски;
- Група 2: задължителна ГО за опасни дейности, асистанс, КАРГО, КАСКО, професионални отговорности.

В таблица 1 са дадени средните стойно-

Table 1 shows gain summary on the decision tree statistics, including the mean values $\bar{Y}$ and the predicted values $\hat{Y}$ of the conversion rate, the number and the percentage of observations for each of the groups in each of the samples (development and validation), as well as the standard deviation.

It could be seen from Table 1 that the validation sample confirms the results from the development sample. In addition, the risk of incorrect classification is low - it has an estimate of 0.019 (or 1.9%) for the development sample and 0.008 (or 0.8%) for the validation sample.

сти $\bar{Y}$ и прогнозните стойности $\hat{Y}$ на нивото на реализация за всяка от групите във всяка една от извадките (работна и валидационна), броят наблюдения и процентното съотношение за всяка от групите в съответната извадка, както и стандартното отклонение. От таблицата става ясно, че валидационната извадка потвърждава резултатите от работната. Освен това, рискът от неправилна класификация е нисък – той има оценка от 0,019 (или 1,9%) за работната извадка и 0,008 (или 0,8%) за валидационната извадка.

Table 1. Decision tree – gain summary by groups
Таблица 1. Дърво на решенията – обобщени статистики по групи

Sample **Извадка**	**Group** **Група**	**# of Cases** **Брой наблюдения**	**% of Sample** **% от извадката**	$\bar{Y}$	$\hat{Y}$	**Std. Dev.** **Станд. откл.**
Development (Training) Работна	1	140	58.6%	0.088	0.101	0.171
	2	99	41.4%	0.055	0.053	0.074
	Total/Общо	239	100.0%	0.074	0.082	0.140
Validation (Test) Валидационна (Тестова)	1	59	59.0%	0.065	0.101	0.079
	2	41	41.0%	0.054	0.053	0.089
	Total/Общо	100	100.0%	0.060	0.082	0.083

Source: *Own Calculations* / **Източник:** *Собствени изчисления*

The distribution of the conversion rate in the two samples has a similar trend (Fig. 2), which confirms the results.

Cross-Validation

Cross-validation divides the sample into a number of subsamples. Decision trees are generated, excluding the data from each subsample in turn: the first tree is generated on all of the records except those in the first subsample; the second tree is generated on all of the records except those in the second subsample, etc. For each of these trees, the stability of the classification is tested using the subsample that was not involved in the tree's generation.

Cross-validation produces a single (final) decision tree.

The risk estimate of incorrect classification for the final tree is the average estimate of the risks for all of the trees.

Разпределението на нивото на реализация в двете извадки има сходна тенденция (Фиг. 2), което затвърждава резултатите.

Крос-валидация

Крос-валидацията разделя извадката на няколко подизвадки. Генерират се дървета на решенията, като се изключват данните от всяка една извадка поотделно: първото дърво се генерира върху всички записи с изключение на тези от първата подизвадка; второто дърво се генерира върху всички записи с изключение на тези от втората подизвадка и т.н. За всяко от тези дървета се прави проверка на стабилността на класификацията, като за целта се използва подизвадката, която не е участвала в генерирането на съответното дърво.

Резултатът от крос-валидацията е едно единствено (финално) дърво на решенията. Валидираната по този метод оценка на риска от неправилна класификация за

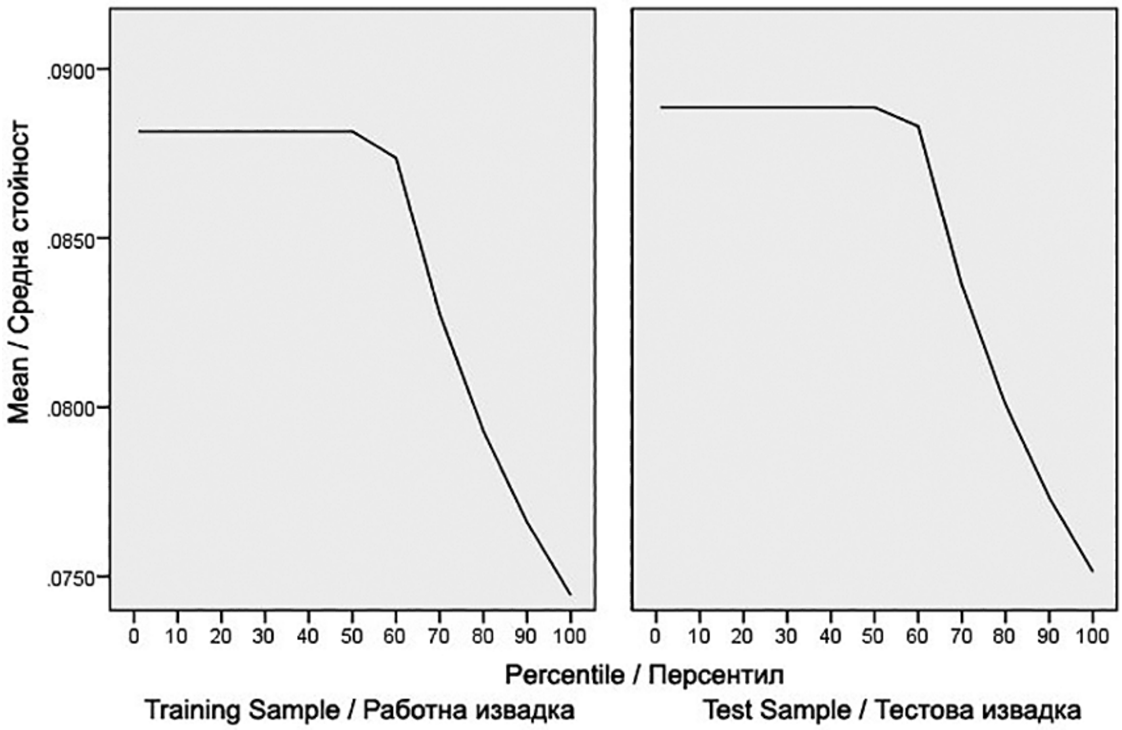

Figure 2. Distribution of the conversion rate in the training and in the test sample

Фигура 2. Разпределение на нивото на реализация в работната и тестовата извадка

Source: *Own Calculations with SPSS /* ***Източник:*** *Собствени изчисления със SPSS*

As a result of the classification, generated on the available data, the observations have been divided into two groups according to the conversion rate of the advertised products (Fig. 3)

- Group 1 (Node 1): Financial Risk Insurance, Real Estate Insurance, Insurance of Construction and Assembly Works, Legal Expenses Insurance
- Group 2 (Node 2): General Liability Insurance, Compulsory Civil Liability Insurance for Dangerous Activities, Travel Assistance Insurance, CARGO, Professional Liability Insurance, CASCO.

It is noteworthy that the grouping with cross-validation confirms the one obtained with the split-sample validation, with one exception: advertising of General Liability Insurance.

Table 2 shows gain summary on the final decision tree statistics. It could be seen from the table that the mean values of the conversion rate for the two groups are very close to those for the respective groups of the development sample in the split-sample validation.

финалното дърво представлява средната оценка на риска за всички дървета.

В резултат от класификацията, генерирана върху наличните данни, наблюденията са разделени в две групи според нивото на реализация на рекламираните продукти (Фиг. 3):

- Група 1: застраховки срещу финансов риск, застраховки на недвижими имоти; застраховки на СМР, застраховане на правни разноски;
- Група 2: обща ГО, задължителна ГО за опасни дейности, асистанс, КАРГО, КАСКО, професионални отговорности.

Групирането тук потвърждава това, получено с разделения подбор, с едно изключение: рекламата на обща ГО.

В табл. 2 са поместени обобщените статистики по групи, откъдето може да се види, че средните стойности на нивото на реализация за двете групи много се доближават до тези за съответните групи от работната извадка при валидирането с разделен подбор.

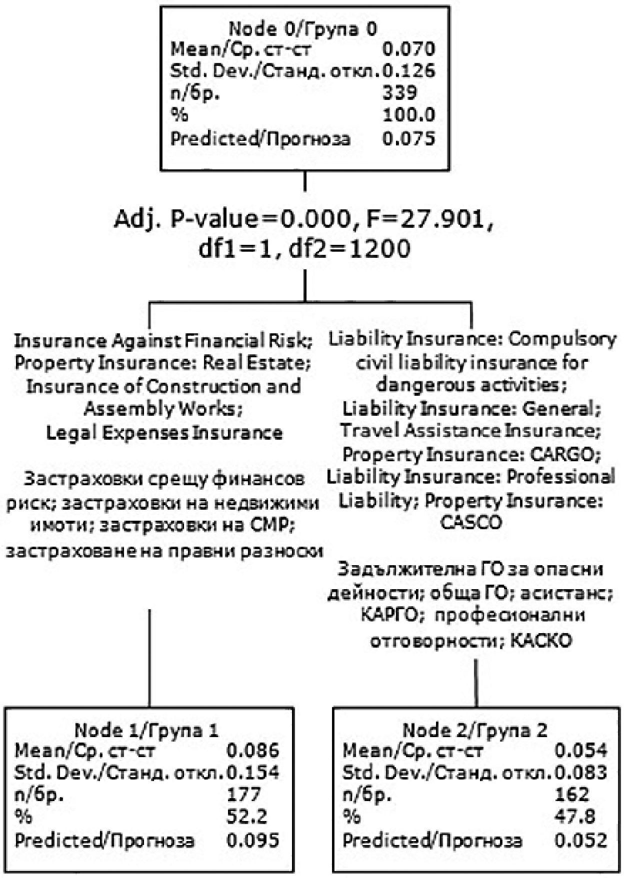

Figure 3. Final decision tree with cross-validation

Фигура 3. Финално дърво на решенията при крос-валидацията

Source: *Own Calculations with SPSS* / ***Източник:*** *Собствени изчисления със SPSS*

Table 2. Final Decision tree – gain summary by groups

Таблица 1. Финално дърво на решенията – обобщени статистики по групи

Group Група	**# of Cases Брой наблюдения**	**% of Sample % от извадката**	$\bar{Y}$	$\hat{Y}$	**Std. Dev. Станд. откл.**
1	177	52.2%	0.086	0.095	0.154
2	162	47.8%	0.054	0.052	0.083
Total/Общо	339	100.0%	0.070	0.075	0.126

Source: *Own Calculations* / **Източник:** *Собствени изчисления*

The risk of incorrect classification is low - it has an estimate of 0.017 (or 1.7%).

The distribution of the conversion rate (Fig. 4) has a similar trend to that of the split-sample validation.

Рискът от неправилна класификация е нисък – той има оценка от 0,017 (или 1,7%).

Разпределението на нивото на реализация (Фиг. 4) има сходна тенденция с това при валидирането с разделен подбор.

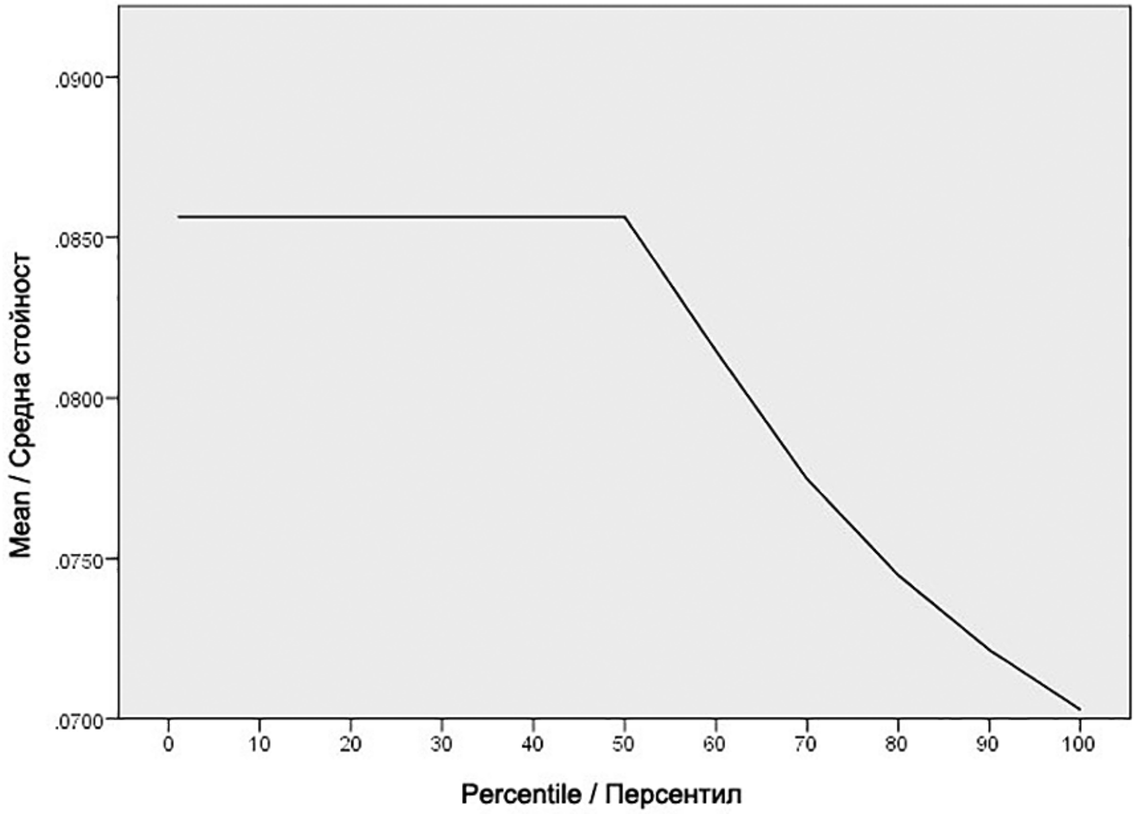

Figure 4. Distribution of the cross-validation conversion rate
Фигура 4. Разпределение на нивото на реализация при крос-валидация

Subsequent Analysis

The results obtained from the two validation methods are similar:

- The cross-validation grouping confirms the one obtained with the split-sample validation with one exception: advertising of General Liability Insurance.
- The mean values of the conversion rate for the two groups obtained through cross-validation are very close to those for the respective groups of the development sample in the split-sample validation.
- The distribution of the cross-validation conversion rate has a similar trend to that of the split-sample validation.

The General Liability Insurance ad has an average conversion rate of 0.045 (or 4.5%), which brings it closer to the mean value for the group with lower conversion rate. Therefore, in the subsequent analysis, the General Liability Insurance ad has been left in Group 2.

Анализ на резултатите

Резултатите, получени по двата метода за валидиране, са сходни:

- групирането при крос-валидацията потвърждава това при разделения подбор с изключение на рекламата на обща ГО;
- средните стойности на нивото на реализация за двете групи, получени чрез кросвалидация, много се доближават до тези за съответните групи от работната извадка при валидирането с разделен подбор;
- разпределението на нивото на реализация при кросвалидацията има сходна тенденция с това при валидирането с разделен подбор.

Рекламата на обща ГО има средно ниво на реализация 0,045, което я доближава повече до средната стойност за групата с по-ниска реализация (отколкото до тази с по-висока реализация). Затова при последващия анализ застраховката обща ГО е оставена в Група 2.

The results obtained from the CHAID procedure can be used to optimize the online advertising strategy of the insurance company. In the interpretation of the results below, the following distribution of the ads to the appropriate groups is respected

- Group 1 (higher conversion rate): Financial Risk Insurance, Real Estate Insurance, Insurance of Construction and Assembly Works, Legal Expenses Insurance
- Group 2 (lower conversion rate): General Liability Insurance, Compulsory Civil Liability Insurance for Dangerous Activities, Travel Assistance Insurance, CARGO, Professional Liability Insurance, CASCO.

This distribution corresponds to the one obtained from the cross-validation method.

Let the company set a minimum conversion rate, below which the ad is considered ineffective. The following scenarios are possible

- Scenario 1: Both groups have an average conversion rate above the minimum - no need for the company to change its strategy.
- Scenario 2: Only Group 1 has an average conversion rate above the minimum - the online advertising strategy for this group may remain as it is, while Group 2 needs a change in the online advertising strategy. If the insurance company wants to reduce its costs, it could directly stop advertising the products from Group 2.
- Scenario 3: Both groups have an average conversion rate below the minimum - then the company could stop the ads of the products from Group 2 and redirect their costs to the products from Group 1 so that Group 1 exceeds the minimum conversion rate.

Получените резултати могат да бъдат използвани за оптимизиране на онлайн рекламната стратегия на застрахователната фирма. При интерпретирането на резултатите по-долу е спазено следното разпределение на рекламите към съответните групи (отговаря на полученото по метода с крос-валидация):

- Група 1: застраховки срещу финансов риск, застраховки на недвижими имоти; застраховки на СМР, застраховане на правни разноски;
- Група 2: обща ГО, задължителна ГО за опасни дейности, асистанс, КАРГО, КАСКО, професионални отговорности.

Нека компанията е заложила минимално ниво на реализация, под което счита рекламата за неефективна. Възможни са следните сценарии:

- Сценарий 1: И двете групи имат средно ниво на реализация над минималното – няма нужда от промяна в стратегията на компанията;
- Сценарий 2: Само Група 1 има средно ниво на реализация над минималното - при нея онлайн рекламната стратегия може да остане, както е, докато Група 2 се нуждае от промяна в онлайн рекламната стратегия. Ако от застрахователната компания искат да намалят разходите си, те биха могли направо да спрат рекламите на продукти от Група 2.
- Сценарий 3: И двете групи имат средно ниво на реализация под минималното – тогава рекламите на продукти от Група 2 могат да бъдат спрени и разходите за тях да бъдат пренасочени към Група 1, така че тя да надскочи минималното ниво на реализация.

III. Conclusion

A product or a service should not only be affordable and of good quality, but also have an effective advertising strategy, in order to achieve good realization on the market. Online advertising services such as Google AdWords provide their customers with statistics on the performance of their ads. Based on this data, each company could optimize its ads in Google's search engine, test new search keywords, stop advertising temporarily, and so on.

To optimize the online advertising strategy, various mathematical tools could be used, one of which – the decision tree. In the work presented, by solving a problem from practice, the potentials of using this tool for classification in the field of online advertising have been outlined.

The classification has been done with SPSS, using the CHAID method. After applying this method to the empirical data, two groups with a significantly different average conversion rate were found. Two different approaches of validating the decision tree have been applied: split-sample validation and cross-validation. They provide reliability about the stability of the classification.

The results of the two validation methods have been compared, and then an analysis has been carried out – which actions the company could take to optimize its online advertising strategy by applying appropriate actions to each of the individual groups (according to their average conversion rate).

The work presented covers a problem that has a practical application in business organizations from different spheres (the field of insurance is used as an example here).

The algorithm presented could be used for other data (regardless of sphere) in a similar situation.

As a next step for future work, the decision tree (or another Data Mining methods) could be considered to optimize some of the other types of online advertising that companies use to reach their customers.

III. Заключение

За да има един продукт или услуга добра реализация на пазара, той трябва не само да е качествен и на добра цена, но и да има ефективна рекламна стратегия. Услугите за онлайн рекламиране като Google AdWords предоставят на клиентите си статистики за представянето на техните реклами. На база на тези данни всяка фирма може да оптимизира рекламите си в търсачката на Google, да изпробва нови ключови думи за търсене, да спре рекламата си временно и т.н.

За оптимизиране на онлайн рекламната стратегия на помощ идват математически инструменти, един от които – дърво на решенията. В настоящата работа чрез решаване на проблем от практиката са изложени възможностите за използване на този инструмент с цел класификация в сферата на онлайн рекламата.

Класификацията е направена с помощта на SPSS, използвайки метода CHAID. Благодарение на метода са намерени две групи със значително различно средно ниво на реализация. Приложени са два различни подхода за валидиране на дървото на решенията: валидиране с разделен подбор и крос-валидация, които дават сигурност относно стабилността на класификацията. Резултатите от двата подхода за валидиране са сравнени, след което е направен анализ на това какви действия може да предприеме фирмата, за да оптимизира рекламната си стратегия, като приложи адекватни действия спрямо всяка една от отделните групи (според средното им ниво на реализация). Настоящата работа обхваща проблем, който има практическо приложение в бизнес организации от различни сфери (за илюстрация е използвана сферата на застраховането). Представеният алгоритъм би могъл да се използва за други данни в сходна ситуация.

Като следваща стъпка за надграждане на работата би могло да се разгледа приложението на дърво на решенията (или друг метод за интелигентен анализ на данни) за оптимизиране на някои от останалите видове онлайн реклама, които фирмите използват, за да достигнат до своите клиенти.

Reference/Литература

Adwords.google.com. (2017). Google PPC Online Advertising | Google AdWords. [online] Available at: https://adwords.google.com/home/ [Accessed 22 Oct. 2017].

Bahari, T. and Elayidom, M. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. Procedia Computer Science, 46, pp.725-731.

Baizyldayeva, U., Uskenbayeva, R. and Amanzholova, S. (2013). Decision Making Procedure: Applications of IBM SPSS Cluster Analysis and Decision Tree. World Applied Sciences Journal, 21(8), 1207-1212.

Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques. Second Edition, The Morgan Kaufmann Series in Data Management Systems.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining Inference and Prediction. Second Edition, Springer. ISBN 978-0-387-84857-0

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. Applied statistics, 119-127.

Kim, J., Lee, B., Shaw, M., Chang, H. and Nelson, M. (2001). Application of decision-tree induction techniques to personalized advertisements on internet storefronts. International Journal of Electronic Commerce, 5(3), pp.45-62.

Linoff, G. and Berry, M. (2011). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Magidson, J. (2005). SI-CHAID 4.0 user's guide. Belmont, MA: Statistical Innovations.

Ngai, E., Xiu, L. and Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), pp.2592-2602.

Olson, D. and Delen, D. (2008). Advanced data mining techniques. Springer Science & Business Media.

Science and Analytics. (2017). IBM SPSS Software | IBM Analytics. [online] Www-01.ibm.com. Available at: http://www-01.ibm.com/software/analytics/spss/ [Accessed 22 Oct. 2017].

Shaw, M., Subramaniam, C., Tan, G. and Welge, M. (2001). Knowledge management and data mining for marketing. Decision support systems, 31(1), pp.127-137.

Song, Y. and Ying, L. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

Support.google.com. (2017). AdWords Help. [online] Available at: http://support.google.com/adwords [Accessed 22 Oct. 2017].

Zaki, M., Meira Jr, W. and Meira, W. (2014). Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.

Abcbg.com. (2017). Интернет реклама и Онлайн маркетинг | ABC Design & Communication. [online] Available at: https://www.abcbg.com/internet-advertising/ [Accessed 27 Nov. 2017].

Гоев, В. (1996). Статистическа обработка и анализ на информацията от социологически, маркетингови и политологически изследвания със SPSS. София.

Иванов, М. (2016). Съвременни методи за интелигентен анализ на данни. Working Paper. Научен електронен архив на НБУ, София.

Манов, А. (2001). Статистика със SPSS. Тракия–М, София.

Павлов, В. & Михова, В. (2016). Приложна статистика със SPSS. АВАНГАРД ПРИНТ, Русе.

Филипов, П. (2014). Приложение на финансови иконометрични техники в моделиране на човешкото щастие. Магистърска теза. Софийски университет „Св. Климент Охридски", София.