Towards use of direct channel for the insurance pricing analytical purposes

Piotr STOLARSKI

Department of Information Systems, Poznań University of Economics and Business

Keywords: Internet, direct insurances, market analysis

# Abstract

Mass use of internet became so far a proven global-scale phenomenon. It is propelled on one hand by improving accessibility and on the other hand by enabling-technologies like broadband. The trend is also visible in traditionally conservative business sectors with finances as key example.

While internet banking has become de facto standard of bank service provision the emerging financial technologies (or fintech) sector have the potential to further transform the business environment as well as introduce innovative business models. The evolution in terms of insurance field took slower pace though. Still it does not mean that the electronic commerce upheaval have not gained its momentum in case of insurances at all. Nowadays not only internet but also the mobile commerce became the carrier of insurance services. Though figures show that these are substantially less intensively chosen as service delivery channel than in case of banking.

In the presented research it has been demonstrated how e-commerce insurance end-points can be employed as the source for acquiring hidden insurance-companies knowledge. Firstly the idea of the process has been explained. Secondly the results of real-life experiments with application of the proposed method are demonstrated. The approach allows for example for an in-depth pricing-policy discovery of insurance products and its further comparison among competitors. The presented results are based on research from [Stolarski2015].

# Introduction

It is estimated that in 2016 47.1% of the global population had access to the internet. It is only little less than half of the planet inhabitants but simultaneously it constitutes also 81.0% of the citizens living in developed countries [ITU2015]. Mass use of internet became a global-scale phenomenon at least in these countries. Yet the developing countries are not much behind. Certainly the percentage of users is much lower in the latter states however they are the driving force of growth in the statistics in recent years with 3.5 percent point growth rate for years 2014-16.

The phenomenon of ubiquitous accessibility is not only reduced to increasing number of users which are able to access services via internet. It is also propelled by the improved quality of the connection as well as development or advancement of services themselves. The good example of the enabling-technologies are broadband where the redefinition of connection speed was necessary to keep up with the progress pace. Another example is the cloud computing.

The trend is also visible in traditionally conservative business sectors especially with the case of finances. As indicated in the EY’ "2015 Global Insurance Outlook" report [Crawford2015] in 2016 any single physical visit to the branch of a financial institution have been accompanied with 250 interactions via the Internet. The "US Online Retail Sales" report by Forester [Mulpuru2011] forecasted that e‑commerce sales in 2010-2015 in Western Europe would grow by 11% to the value of 115 billion euros. This potentially could influence growth in the insurance sector as well.

While internet banking has become de facto standard of bank service provision the emerging financial technologies sector have the potential to further change the business environment as well as introduce innovative business models. According to Eurostat in 2016 the share of individuals using e-banking in Poland equaled 39% of the population aged 16 to 74. And it raised by 25% comparing to the year earlier. The average for EU27 was 49% [Eurostat2017].

The evolution in terms of insurance field took slower pace. While the average of individuals having at least one transaction of buying or renewing an insurance product via internet for EU27 amounts to 11% in 2016 there are vast differences among individual states. In the Netherlands it is up to 31% while only 1% in countries like Bulgaria or Romania. The impact of the emergence of this sale channel on the functioning of insurance companies has been discussed among others in [Kaczała2006].

It does not mean that the electronic commerce revolution have not gained its momentum in case of insurances. Insurance sales on-line becomes increasingly popular [Forrester2011, PIU2011]. Statistics given by McKinsey & Company [Junker2014] indicate that global sales of direct insurance increased from $ 100 billion in 2008 to $ 129 billion by 2012.

Nowadays not only internet but also the mobile commerce became the carrier of financial services. It is still the technology of the future for insurance though mobile banking becomes the next wave of the banking sector. In Poland at the end of 2016 there were about 7.7 million active mobile banking solutions users [PRNews2017].

In the presented research it has been demonstrated how e-commerce insurance end-points can be employed as the source for acquiring internal insurance-companies knowledge. The insurance knowledge web sources like premium calculators can be used to acquire knowledge on the way those premiums are computed. The gained knowledge allows comprehensive pricing-policy discovery.

In order to be able to gather the knowledge and build pricing models in a systematic manner a method have been prepared. One can enumerate a number of applications of the elaborated method. The most promising are: insurance market monitoring, supplying data for offer aggregating portals, supporting an alternative interoperability model, and scientific research purposes.

The paper is divided into several sections. In the section 1 the important works associated with the topic of the research are summarized. The method of research and basic facts on data gathered during study are given in section 2. Section 3 describes the process of model generation out of datasets. Moreover the section provides a preliminary evaluation of those models. A case study which accompanied the research is presented in section 4. Section 5 is devoted to rankings of analytical methods being in use for generating premium models while section 6 is the summary of the work carried out.

# Related works

## Data Mining

Data mining is a field of discovery, study and application of broad range of common analytical methods and techniques to large datasets with the intention of detection of unknown patterns. Those methods encompasses data processing procedures of diversified nature and origin, including more traditional techniques coming from statistics like regression analysis, cluster analysis, moving to more-AI related, i.e. artificial neural networks, decision rules or genetic algorithms.

In order to obtain meaningful results as well as sustain quality of data processing the application of data mining methods should be executed within a framework. The framework defines a context and processing stages in the shape of a process-like sequence. The examples of such a frameworks are CRISP-DM or SEMMA to name only few [Kurgan2006].

Data mining techniques are widely operated in a number of fields, esp.: business and marketing, economics as well as insurance.

## Information extraction

[Chang2006] defines Information Extraction (IE) from web sources as “automated transformation of WWW documents into structured data”. The extraction process takes un- or weakly structured information as the input, whereas it outputs the information with entirely fixed structure. According to [Iskold2007] when it comes to classify the approaches to web sources information structuring, one can identify two main trends: the bottom-up and the top-down.

In the bottom-up approach it is the creators of the web pages content who are responsible for proper tagging and annotating of the text so that it is easily automatically processed. The content creators if willing may use a number of well-developed technologies or formal languages like: XML, XSLT [Clarck1999], RSS, RDF [Beckett2004], RDFa, OWL [McGuinness2004], JSON, and DublinCore. Other so called microformats can also be utilized to realize the aim of the described approach.

The top-down approach represents completely different philosophy. In this case the way in which information is being publicized in the web remains (almost) unchanged. Instead algorithms which aim at discovery and acquisition of information are commonly used in order to tackle the unknown web content structure problem. It means that some special systems and software is needed in order to realize this approach. The software is then responsible for the information extraction tasks.

It has to be noted that due to various reasons the top-down trend is more real-life case. The mentioned IE systems are grouped by [Hsu1998] into: manually created shells programmed in common programming languages, manually created shells with special language, heuristic-driven shells and induction shells. As a contrast [Kushmerick2003] among IE systems distinguishes: finite-state systems and systems using relational learning.

## Web mining

Web mining is the “use of data mining techniques to automatically discover and extract information from Web documents and services” [Konopnicki1998]. Typically there are three main areas of web mining with distinguishable differences of purpose and processing source type. These are: Web content mining, Web structure mining and Web usage mining [Madria1999].

Web content mining is the most active field o the three research areas. It is focused on obtaining better structured or completely new information from a wide variety of WWW documents. It should be noted that the information extraction mentioned earlier is considered to be vital part of Web content mining. However there is a vast number of other examples of works done within this field. They include such tasks as: categorization / clustering or text summarization. More advanced tasks aims at named entity and relation detection, schema discovery up to ontology learning. With the emerging of social web and the explosive growth of web 2.0 content, works related to such tasks as: opinion / sentiment analysis (mining), user tags knowledge discovery become an important part of the researches as well.

Web structure mining is the study of the topology of the hyperlinks graph between WWW documents. The field focuses preliminarily on algorithms for measuring the significance of Web sites within the interrelated graph of links. This type of algorithms have their practical use, for instance, in search engines with PageRank being a good example [Page1998]. Besides this main point of interest Web structure mining includes also community discovery Web site complexity measurement and Web page categorization.

Finally, Web usage mining is defined by [Kolari2004] as “analysis of results of user interactions with a Web server, including Web logs, clickstreams, and database transactions at a Web site or a group of related sites”. The focus here is at discovery of interesting behavioral patterns of browser users both individually and at the collective level. According to [Srivastava2000] Web usage mining uses three sources of information: server-level data, data collected at client-level and traffic data captured by proxies. The results of Web usage mining algorithms may be used by business intelligence systems but also foster the improvement of web sites structure. Other uses include automated recommendations or content / navigation personalization.

## Knowledge discovery from web sources

With our research we argue that by combining data mining methods with information extraction it is feasible to expand web mining ideas. This expansion means that given a well-specified domain selected web sources could be treated as knowledge discovery sources.

The differences between Deep web which is part of advanced information extraction approach, web mining and web knowledge discovery is summarized on Figure 1.

Figure 1. Deep web, web mining and knowledge extraction – comparison of approaches

Deep web

Web knowledge discovery

|  |  |  |  |
| --- | --- | --- | --- |
|   |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Web mining

www

www

Web service

www

www

www

classification

In the Deep web the database is the actual source. The web infrastructure in this case is merely an intermediary interface for data access. Consequently, the number of records and possible queries is strictly limited. Web mining as described in the section above is in fact a diversified field where often metadata plays crucial role in execution of some common tasks like classification. Finally in case of knowledge discovery from web sources the source is a model (an algorithm). There is no single database which has a number of consequences like esp. substantially vaster space of queries.

## Ratemaking and actuarial techniques

The conducted research is mainly concerned with calculation of premium levels and pricing schemes of differentiated group of insurance products. Typically these kind of settings are the result of diversified actuarial procedures. That is why the presented research is closely related to actuarial knowledge [Ronka-Chmielowiec2003].

The relation is twofold. On one hand the knowledge of ratemaking and actuary is helpful in order to acquire sound data and process them into meaningful models. On the other hand the models produced are to simulate the genuine models. That is why the results of this study may let deduce on ratemaking practice in large picture.

A very thorough yet basic elaboration of commonly used ratemaking and actuarial techniques is given in [Werner2010]. Nevertheless a large number of less common approaches can be pointed out. This include for instance the use of data mining and analytical tools [Hastie2009] or simulation [Salam2003].

# Data and method of the empirical study

## The method of insurance premium models extraction from web sources

The method described in the paper may be considered to be the part of web mining researches as presented in Related Works section. If so it should be placed within the category of web content mining due to its resemblance on the conceptual level as well as some common characteristics as stated in Knowledge discovery from web sources section. It is also in line with the definition of knowledge extraction presented in [Kosala2000]. Yet the method and the research is rooted in the specific domain of insurances and as such there are no similar solutions reported to the presented method and task of premium model extraction.

The method is divided into two separate phases: the preparatory phase and the executive phase. Three main components have been designed in order to support the implementation of the stages contained in both phases.

The preparatory phase includes the following stages: testing sources, declaration of candidates for the parameters of the model, construction of a navigation graph, refinement of navigation graph, control concepts annotation and model test. The navigation planner component is being used in all above stages but the last one. In the model test stage the method projects the usage of web source representation explorer component.

The executive phase consists of stages: data extraction, iteration on property values set, optimization strategies selection, alternative models creation and – finally - choosing the solution. The method assumes the use of navigation and extraction components as well as external analytical tools to support this phase.

The presented method employs external analytical tools in order to build the ultimate model candidates. Every candidate model relates level of premium to diversified factors reflected by rating or underwriting variables. In the conducted study, 22 models have been applied for each dataset using any of the tested tools. Half of them have been build using the original data and another half have been developed based on data transformed by principal component analysis (PCA). The tested analytical tools included: artificial neural networks [Rumelhart1986], automated neural networks, decision trees [Quinlan1996], regression, DMINE procedure, partial least squares method, memory-based methods, boosting trees, least-angle regression (LARS).

Table 1 Information about the analytical methods (techniques) used to construct models

|  |  |  |
| --- | --- | --- |
| Method symbol | Method | Comment |
| Neural | artificial neural networks (ANN) |  |
| Ensmbl | aggregation | Model being the aggregation of results of other non-PCA models |
| AutoNeural | automated neural networks | Automatically generated ANN |
| Tree | decision trees |  |
| DmineReg | DMINE procedure |  |
| PLS | partial least squares method |  |
| Reg | classic regression |  |
| DMNeural | DMNeural |  |
| LARS | least-angle regression (LARS). |  |
| MBR-PCA | memory-based methods | Model using data after PCA transformation |
| Boost | boosting trees |  |
| Tree-PCA | decision tree | Model using data after PCA transformation |
| Boost-PCA | boosting trees | Model using data after PCA transformation |
| Ensmbl-PCA | aggregation | Model being the aggregation of results of other PCA models |
| DmineReg-PCA | DMINE procedure | Model using data after PCA transformation |
| AutoNeural-PCA | automated neural networks | Model using data after PCA transformation |
| Neural-PCA | artificial neural networks | Model using data after PCA transformation |
| PLS-PCA | partial least squares method | Model using data after PCA transformation |
| Reg-PCA | classic regression | Model using data after PCA transformation |
| DMNeural-PCA | DMNeural | Model using data after PCA transformation |
| LARS-PCA | LARS | Model using data after PCA transformation |
| MBR | memory-based methods |  |

All the methods used in the study are enlisted in Table 1. The symbols given in the table will be referred further for the purposes of evaluation results presentation. The labels for models based on data transformed by principal component analysis are suffixed with “PCA”.

## The data sets

In the empirical study carried out, data from 19 insurance web sources, belonging to 13 insurance companies have been acquired. The data were gathered between May 2015 and June 2016. The data were further divided according to the criterion of the type of risk. It was particularly important in the case of complex insurance products. For the sake of the research the complex insurance products were defined as insurances which either are protecting against distinctive but usually related risks or are effectively a cross-sale compound. In both cases the product was considered complex and hence the data division took place only if the character of data and the product offer allowed for certain identification of the complex nature. Such case of certain complexity identification is for instance when listed sub items are pricing or there exist an explicit possibility to decline the optional risk protection.

The employed method and tools prepared for its execution is language and culture independent. In the reported study only polish web sites of the insurers were taken into account as potential data and knowledge sources. This was dictated by the pioneering application of the method and for the sake of expected grater consistency of results.

On the other hand a rather broad perspective of insurance products types were taken into account ranging from property to life insurances. Those product types are summarized in Table 2.

The table also demonstrates the amount of raw data obtained during the initial phase of the knowledge discovery method, namely the information extraction stage. Typically a minimal number of extracted information was set to 2000 records per web source. The setting was adjusted on case by case basis regarding the (expected) complexity of the premium computation model. The complexity was derived primarily from the number of independent variables.

Table 2 The number of raw data records collected in the extraction process coupled with the source addresses and division into insurance types

|  |  |  |  |
| --- | --- | --- | --- |
| Source | Address | Type | Number of records |
| 1 | allianzdirect.pl | motor | 3061 |
| 2 | aviva.pl | travel | 2129 |
| 3 | axadirect.pl | motor | 2382 |
| 4 | ehome.benefia24.pl | home | 2069 |
| 5 | emoto.benefia24.pl | motor | 2393 |
| 6 | kuke.com.pl | commerce | 2723 |
| 7 | libertydirect.pl (lu.pl) | motor | 2607 |
| 8 | link4.pl | motor | 2572 |
| 9 | mtusa.pl | motor | 2505 |
| 10 | signal-iduna.pl | life | 1202 |
| 11 | skokubezpieczenia24.pl | home | 1842 |
| 12 | skokubezpieczenia24.pl | life | 1091 |
| 13 | skokubezpieczenia24.pl | motor | 2800 |
| 14 | tutum.pl | bicycle | 365 |
| 15 | tutum.pl | against accidents | 1728 |
| 16 | uniqa24.pl | home | 2492 |
| 17 | youcandrive.pl | travel | 2017 |
| 18 | youcandrive.pl | home | 2136 |
| 19 | youcandrive.pl | motor | 1994 |
|  |  |  |  |
|  |  | **TOTAL** | 40108 |

As stated in the method description the information extraction from web sources is the intermediary step of insurance knowledge discovery. It is an important stage because the gathered information will feed the subsequent phases of the method. This stage may be characterized by its own measure of efficiency which is the percentage of valid results returned by the web source. The valid result is not the same as valid single web server response. It is because it takes usually several interactions with web application calculating the premium in order to obtain a proper premium quotation.

Table 3 shows the indicators of information extraction step efficiency per each web source. The invalid part comprise both the communication errors and any other situations when the premium could not been calculated. One of the reasons of inability of calculating the premium was to fail to qualify for underwriting due to the submitted values of underwriting variables. The average value of valid indicator amounting more than 93% should be considered as very high yet the efficiency of the extraction process for certain web sources differ adversely. Usually it was the case of less usual or more complicated - derived – products.

Table 3 The share of valid and invalid records obtained in the extraction process, specifying the data sources

|  |  |  |
| --- | --- | --- |
| Source | % valid | % invalid |
| allianz#1 | 100,00% | 0,00% |
| allianz#2 | 100,00% | 0,00% |
| aviva#1 | 98,00% | 2,00% |
| axadirect#1 | 86,39% | 13,61% |
| axadirect#2 | 77,57% | 22,43% |
| ehome.benefia24#1 | 95,28% | 4,72% |
| emoto.benefia24#1 | 98,83% | 1,17% |
| kuke#1 | 99,30% | 0,70% |
| libertydirect#1 | 92,77% | 7,23% |
| libertydirect#2 | 98,72% | 1,28% |
| link4#1 | 92,00% | 8,00% |
| link4#2 | 92,03% | 7,97% |
| mtusa#1 | 99,70% | 0,30% |
| signal-iduna#1 | 94,84% | 5,16% |
| skokubezpieczenia24.health#1 | 70,33% | 29,67% |
| skokubezpieczenia24.home#1 | 99,27% | 0,73% |
| skokubezpieczenia24.home#2 | 96,90% | 3,10% |
| skokubezpieczenia24.moto#1 | 98,94% | 1,06% |
| skokubezpieczenia24.moto#2 | 91,62% | 8,38% |
| skokubezpieczenia24.moto#3 | 91,59% | 8,41% |
| tutum.bike#1 | 98,93% | 1,07% |
| tutum.nnw#1 | 76,86% | 23,14% |
| uniqa24#1 | 96,07% | 3,93% |
| uniqa24#2 | 96,07% | 3,93% |
| youcandrive.moto#1 | 99,40% | 0,60% |
| youcandrive.moto#2 | 96,08% | 3,92% |
| youcandrive.moto#3 | 96,68% | 3,32% |
| youcandrive.travel#1 | 95,39% | 4,61% |
| youcandrive.travel#2 | 80,06% | 19,94% |
|  |  |  |
| AVERAGE | 93,44% | 6,56% |

# Model generation and estimation

As mentioned earlier any candidate model relates premium as dependent variable to a set of independent variables representing particular factors reflected by rating or underwriting variables. It should be stressed that every web source is connected to its own unique internal rating model and as a result every data set associated with a given web source encompasses its own set of independent variables.

Table 4 Comparison of the number of independent variables according to variable types by data sets

|  |  |  |  |
| --- | --- | --- | --- |
| Data set | Temporal | Interval | Nominal |
| **Binary** | **Multivariate** |
| allianz#1 | 2 | 8 | 4 | 8 |
| allianz#2 | 2 | 8 | 4 | 8 |
| aviva#1 | 2 | 13 | 3 | 1 |
| axadirect#1 | 2 | 38 | 7 | 39 |
| axadirect#2 | 2 | 39 | 7 | 39 |
| ehome.benefia24#1 | 0 | 21 | 4 | 1 |
| emoto.benefia24#1 | 0 | 13 | 3 | 7 |
| kuke1#1.com | 0 | 3 | 0 | 1 |
| libertydirect#1 | 1 | 35 | 5 | 6 |
| libertydirect#2 | 1 | 34 | 5 | 6 |
| link4#1 | 2 | 36 | 7 | 25 |
| link4#2 | 2 | 37 | 7 | 25 |
| mtusa#1 | 1 | 9 | 3 | 7 |
| signal-iduna#1 | 1 | 10 | 2 | 16 |
| skokubezpieczenia24.health#1 | 2 | 16 | 3 | 2 |
| skokubezpieczenia24.home#1 | 2 | 39 | 2 | 2 |
| skokubezpieczenia24.home#2 | 2 | 39 | 2 | 2 |
| skokubezpieczenia24.moto#1 | 1 | 42 | 4 | 11 |
| skokubezpieczenia24.moto#2 | 1 | 43 | 4 | 11 |
| skokubezpieczenia24.moto#3 | 1 | 43 | 4 | 11 |
| tutum.bike#1 | 3 | 8 | 2 | 2 |
| tutum.nnw#1 | 3 | 9 | 2 | 2 |
| uniqa24#1 | 0 | 25 | 4 | 2 |
| uniqa24#2 | 0 | 26 | 4 | 2 |
| youcandrive.home#1 | 1 | 14 | 12 | 4 |
| youcandrive.moto#1 | 2 | 17 | 3 | 12 |
| youcandrive.moto#2 | 2 | 17 | 3 | 12 |
| youcandrive.moto#3 | 2 | 17 | 3 | 12 |
| youcandrive.travel#1 | 3 | 8 | 16 | 4 |
| youcandrive.travel#2 | 3 | 8 | 16 | 4 |
|  |  |  |  |  |
| AVERAGE | 1,534 | 25,5 | 5,13 | 10,65 |
| SHARE | 3,59% | 59,54% | 11,99% | 24,88% |

For the sake of data processing and subsequent model generation the variables representing particular factors reflected by rating or underwriting variables were categorized into several types. The categorization and quantitative summary are given in Table 4.

The models have been evaluated using a commonly used measures - primarily mean square error (MSE) and alternatively coefficient of determination (R-square). The mean square error measure values have been enlisted for each best model generated for every dataset in Table 5. The MSE training value is for the training subset of each dataset while the MSE validation value describes the amount of error achieved for validation data subset. The training subsets were created by random selection of 40% of datasets records.

Table 5 Measures reflecting the quality of the computed models

|  |  |  |  |
| --- | --- | --- | --- |
| Data set | Best model | MSE training | MSE validation |
| allianz#1 | Neural | 715929,04 | 927281,52 |
| allianz#2 | DmineReg | 253638,19 | 203726,55 |
| aviva#1 | AutoNeural-PCA | 8,216 | 13,876 |
| axadirect#1 | Ensmbl | 1059,65 | 1076,41 |
| axadirect#2 | AutoNeural | 8452,55 | 9576,6 |
| ehome.benefia24#1 | Neural-PCA | 397,44 | 360,3 |
| emoto.benefia24#1 | Tree | 5844,64 | 9070,1 |
| kuke#1 | Tree | 339822193,7 | 428866297,5 |
| libertydirect#1 | Tree | 68303,87 | 64766,32 |
| libertydirect#2 | Neural | 2182,52 | 2643,51 |
| link4#1 | Neural | 11656,16 | 9619,67 |
| link4#2 | Neural | 45722,63 | 47853,89 |
| mtusa#1 | Neural | 9583,68 | 10099,51 |
| signal-iduna#1 | Tree | 16029,1 | 35439,85 |
| skokubezpieczenia24.health#1 | Tree | 66,32 | 66,65 |
| skokubezpieczenia24.home#1 | Tree | 887,99 | 914,65 |
| skokubezpieczenia24.home#2 | LARS | 87,56 | 122,34 |
| skokubezpieczenia24.moto#1 | Neural | 21015,03 | 25713,57 |
| skokubezpieczenia24.moto#2 | AutoNeural | 84650,13 | 113127,6 |
| skokubezpieczenia24.moto#3 | Tree | 95068,83 | 190134,43 |
| tutum.bike#1 | Tree | 1,82 | 2,41 |
| tutum.nnw#1 | AutoNeural | 1907,95 | 2810,66 |
| uniqa24#1 | Tree | 2,29 | 1,98 |
| uniqa24#2 | MBR | 2494,38 | 3066,18 |
| youcandrive.home#1 | LARS | 65,64 | 57,44 |
| youcandrive.travel#1 | Tree | 114314,16 | 120255,59 |
| youcandrive.travel#2 | Tree | 6821,32 | 7575,29 |
| youcandrive.moto#1 | Neural | 65844,68 | 72560,32 |
| youcandrive.moto#2 | DmineReg | 11533,53 | 16371,72 |
| youcandrive.moto#3 | DmineReg | 15369,85 | 20208,04 |

# The case of pricing-policy drill-down

As an additional study a comparison of pricing models of two product offers were carried out. The compared products reflected information acquired from the mtsu.pl and skokubezpieczenia24.pl web sources (datasets: mtusa#1 and skokubezpieczenia24.moto#1). Both products constitute example of motor insurance.

A careful observer will notice the initial difference in both sources just by looking at the list of web sources in general as presented in Table 5. It can be seen that mtsu.pl offered only single plain product whereas skokubezpieczenia24.pl had a complex one protecting against a number of risks. It is why there are 3 datasets named skokubezpieczenia24.moto.

In general any pricing model in order to compute the individual premium have to input the description of object of insurance. A comparison of any two premium models require a mechanism of submitting equivalent descriptions of the same object of insurance. Therefore the creation of special dictionaries describing the object of insurance which was a limited subset of cars (multilevel classifications) had to be performed. It is a common practice of insurers to use some standardized third-party classifications or entire databases like [EuroTax2017]. In case of the compared models both of them had similar schema of car specification. The pattern for mtsu.pl was: brand → vintage → engine type → capacity → gear type → number of doors → model → type. The stage of dictionary creation was realized semi-automatically with the use of prepared automation scripts.

The next stage in comparison is usually matching the dictionary entries. In order to achieve this the dictionaries are transformed into common ontology. The transformation process is typically manual. Thanks to the placement of items of both dictionaries in one knowledge structure it is possible to automatically translate entries from one classification (dictionary) to the entries of the other classification. In the case of the reported comparison the step was mainly formal procedure because of mentioned schema resemblance. For example a following string of values: “354; 2005; D; 1968; M; 5; 11836; 53964” represents a potential insurance object with subsequent properties: brand = AUDI, vintage = 2005, engine = diesel, capacity = 1968, gear = manual, doors = 5, model = A4 Quattro 2.0 TDi Kat. 8E, type = DPF both in the mtsu.pl web source and in skokubezpieczenia24.pl web source.

It has to be noted that because both classifications were very similar for those specific sources the process of conversion to common ontology was straightforward. In case of most of other sources this favorable circumstances may not take place. Thus the uniformity conversion process would be much more difficult for corresponding structures.

The model comparison revealed that the main differentiating pricing factor for both products is the car location characterized by registration area. The detected differences are given in Table 6.

The results shows that the geography of detected differences is based on the denser structure than city itself. For instance in case of Warsaw (Warszawa) city the differences vary internally.

Table 6 The detected differences between the levels of premiums in comparison of models based on mtusa.pl and skokubezpieczenia24.pl sources

|  |  |  |
| --- | --- | --- |
| Area | Location index (postal code) | Premium difference |
| Bełchatów | 97-400 | 4,96% |
| Biała Podlaska | 21-500 | 5,01% |
| Bydgoszcz | 85-455 | -0,24% |
| Chorzów | 41-519 | 7,13% |
| Dębica | 39-202 | 7,13% |
| Elbląg | 82-300 | 3,98% |
| Gliwice | 44-100 | 4,95% |
| Gniezno | 62-200 | 5,02% |
| Jastrzębie-Zdrój | 44-335 | 1,88% |
| Kalisz | 62-800 | 4,95% |
| Kołobrzeg | 78-100 | 4,99% |
| Konin | 62-500 | 5,01% |
| Kraków | 30-109 | -0,24% |
| Krosno | 38-400 | 10,27% |
| Łódź | 94-020 | 5,04% |
| Lubin | 59-300 | 1,84% |
| Lubin | 59-304 | 1,84% |
| Lublin | 20-105 | 3,74% |
| Mysłowice | 41-401 | 7,09% |
| Nysa | 48-300 | 2,59% |
| Opole | 45-076 | 4,94% |
| Ostrów Wielkopolski | 63-400 | 2,20% |
| Piaseczno | 05-500 | 3,30% |
| Piotrków Trybunalski | 97-304 | 5,00% |
| Poznań | 60-660 | 3,30% |
| Poznań | 61-530 | 3,30% |
| Pruszków | 05-809 | 2,57% |
| Przemyśl | 37-707 | 7,15% |
| Radomsko | 97-500 | 2,58% |
| Rybnik | 44-200 | -0,24% |
| Siemianowice Śląskie | 41-100 | 5,02% |
| Skarżysko-Kamienna | 26-110 | 3,12% |
| Sosnowiec | 41-219 | 7,06% |
| Stargard Szczeciński | 73-100 | 2,57% |
| Świdnica | 58-100 | 2,58% |
| Szczecin | 70-840 | 2,57% |
| Tarnobrzeg | 39-400 | 2,58% |
| Tczew | 83-101 | 2,59% |
| Tychy | 43-102 | 3,49% |
| Warszawa | 03-112 | 2,93% |
| Warszawa | 02-783 | 2,93% |
| Wejherowo | 84-200 | 4,37% |
| Wręczyca Mała | 42-130 | 6,35% |
| Wrocław | 50-065 | 9,72% |

# Determination of best-performing analytical methods for pricing model discovery

In order to identify the analytical tools which produce best results a number of rankings have been prepared. The idea is to detect the best analytical tools for the task of building of premium pricing insurance models. The rankings are based on MSE measures. Table 7-9 present the position of analytical tools according to different ranking criteria.

Table 7 Ranking of all analytical methods according to the number of best fitted (first rank) models

|  |  |  |
| --- | --- | --- |
| # | Analytical method | Number of best models |
| 1 | Tree | 11 |
| 2 | Neural | 7 |
| 3 | AutoNeural | 3 |
| 4 | DmineReg | 3 |
| 5 | LARS | 2 |
| 6 | AutoNeural-PCA | 1 |
| 7 | Ensmbl  | 1 |
| 8 | MBR | 1 |
| 9 | Neural-PCA | 1 |
| 10 | Boost | 0 |
| 11 | Boost-PCA | 0 |
| 12 | DmineReg-PCA | 0 |
| 13 | DMNeural | 0 |
| 14 | DMNeural-PCA | 0 |
| 15 | Ensmbl-PCA | 0 |
| 16 | LARS-PCA | 0 |
| 17 | MBR-PCA | 0 |
| 18 | PLS | 0 |
| 19 | PLS-PCA | 0 |
| 20 | Reg | 0 |
| 21 | Reg-PCA | 0 |
| 22 | Tree-PCA | 0 |

The evaluation of the set of prepared models shows that the decision trees is the data mining technique best suited to produce derived premium models (12 best models). One can draw a conclusion that for the purpose of insurance premium modelling this way of model reconstruction significantly outperform the rest of analytical techniques presented in the paper.

Other two rankings from Table 8 and 9 reveals the position of individual analytical tools according to the accumulated amount of normalized errors which those tools gained for each model produced with the use of particular analytical tool. The lower the sum of errors the better the models produced overall. While Table 8 gives the figures of errors calculated for validation subset, the successive table summarizes the figures for training subset. Obviously, the validation subset is more important as it reflects better the real prediction capability of the models.

Table 8 Ranking of all analytical methods according to the normalized sum of errors for model validation subsets

|  |  |  |
| --- | --- | --- |
| # | Analytical method | Normalized errors sum (validation) |
| 1 | Tree | 13,90587883 |
| 2 | Neural | 18,0497824 |
| 3 | Ensmbl | 18,16807449 |
| 4 | DmineReg | 18,78294511 |
| 5 | LARS | 22,17795715 |
| 6 | Reg | 24,23950957 |
| 7 | PLS | 25,62420136 |
| 8 | DMNeural | 27,23978284 |
| 9 | Boost | 27,76531549 |
| 10 | MBR-PCA | 30,47225519 |
| 11 | Neural-PCA | 31,34130525 |
| 12 | Tree-PCA | 31,34917927 |
| 13 | AutoNeural | 31,37106041 |
| 14 | Ensmbl-PCA | 32,08284908 |
| 15 | MBR | 32,46942853 |
| 16 | DmineReg-PCA | 33,33598569 |
| 17 | Boost-PCA | 35,66328588 |
| 18 | Reg-PCA | 36,79349825 |
| 19 | LARS-PCA | 37,0883756 |
| 20 | PLS-PCA | 37,24142393 |
| 21 | DMNeural-PCA | 37,61626668 |
| 22 | AutoNeural-PCA | 44,96137412 |

In the case of data presented in Table 8 and 9 the relatively small differences of accumulated amount of normalized errors (esp. those corresponding to the top positions of rankings) allow to conclude that the models have quite high prediction capabilities as they preform not worse when working on earlier unknown data than when working on data used to create and train the models.

# Summary

The paper presents the results of the experiment of discovering knowledge on premium calculation. As more and more insurers offer a direct access to the functionality of computing an individual premium for the offered insurance service the internet distribution channel have the specific property of being treated as knowledge source.

Web sources with insurance calculators are used in order to discover the way in which the calculation process is executed. The experiment was conducted in a systematic manner according to the elaborated method.

In the study the model of calculations were discovered for 19 web sources representing different insurance products. 22 variants of diversified analytical tools were tested to generate premium calculation models candidates for each dataset provided by every web source.

One of possible application of the method is the comparison of the models for instance of competing products. The paper shows a simple instance of such comparison as use case.

Table 9 Ranking of all analytical methods according to the normalized sum of errors for model training subsets

|  |  |  |
| --- | --- | --- |
| # | Analytical method | Normalized errors sum (training) |
| 1 | Tree | 12,79551593 |
| 2 | Neural | 16,59044649 |
| 3 | Ensmbl | 16,99692072 |
| 4 | DmineReg | 17,54329682 |
| 5 | LARS | 21,22148459 |
| 6 | Reg | 22,11998726 |
| 7 | PLS | 24,31110898 |
| 8 | DMNeural | 25,94245684 |
| 9 | Boost | 26,0086734 |
| 10 | MBR-PCA | 27,4790531 |
| 11 | Tree-PCA | 27,54750711 |
| 12 | DmineReg-PCA | 27,66539588 |
| 13 | Neural-PCA | 28,68228791 |
| 14 | MBR | 29,75349386 |
| 15 | Ensmbl-PCA | 29,84102391 |
| 16 | AutoNeural | 29,98527432 |
| 17 | Boost-PCA | 33,38611637 |
| 18 | Reg-PCA | 34,90339267 |
| 19 | LARS-PCA | 34,92601823 |
| 20 | DMNeural-PCA | 35,3419063 |
| 21 | PLS-PCA | 35,53701051 |
| 22 | AutoNeural-PCA | 43,36476126 |

In general 418 instances of premium calculation models were prepared during the study. A number of rankings of analytical tools usefulness for best fitted modelling were created with the use of this model population. The rankings indicate that the preferred analytical tool for discovering premium models are decision trees.

The important point of the study is that the method is language and culture agnostic as well as it is largely independent of the insurance product type.

# References

[Beckett2004] D. Beckett, B. McBride, *RDF/XML syntax speciﬁcation*, 2004.

[Chang2006] Ch. Chang, M. Kayed, M. Ramzy Girgis, K. Shaalan. *A survey of web information extraction systems*. IEEE Transactions on Knowledge and Data Engineering, 18(10):1411–1428, 2006.

[Crawford2015] S. Crawford, et al., *2015 Global insurance outlook*, E&Y, 2015.

[Eurostat2017] *Individuals using the internet for internet banking*. Eurostat 2017. Retrieved: 2017-06-01.

[EuroTax2017] <https://www.autovistagroup.com/>. Retrieved: 2017-06-01.

[Forrester2011] *US Online Insurance Forecast, 2010 To 2015*, Forrester Research, Inc. 2011.

[Hastie2009] T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Min-ing, Inference and Prediction*, Springer Science & Business Media, ISBN 9780387848587, 2009.

[Hsu1998] Ch. Hsu, M. Dung, *Generating ﬁnite-state transducers for semi-structured data extraction from the web*, Information Systems, 23(9), s. 521–538, 1998.

[Iskold2007] A. Iskold*, Top-down: A new approach to the semantic web*, 2007.

[ITU2015] *ICT Facts and Figures 2005, 2010, 2016*. Telecommunication Development Bureau, International Telecommunication Union (ITU). <http://www.itu.int/en/mediacentre/pages/2016-PR30.aspx>. Retrieved: 2017-06-01.

[Junker2014] L. Junker, S. Gerssen, M. Jutte, *Global Insurance Industry Insights: An in-depth perspective*, McKinsey & Company, 2014.

[Kaczała2006] M. Kaczała, *Internet jako instrument dystrybucji ubezpieczeniowej*, PhD Thesis, UEP 2006.

[Kolari2004] P. Kolari, A. Joshi. *Web mining: Research and practice*. Computing in Science and Engineering, 6(4):49–53, 2004.

[Konopnicki1998] D. Konopnicki, O. Shmueli. *W3QS: A query system for the world-wide web*. In 21st International Conference on Very Large Data Bases, 1995.

[Kosala2000] R. Kosala, H. Blockeel, *Web mining research: a survey*. ACM SIGKDD Explorations Newsletter, 2(1):1–15, 2000.

[Kurgan2006] L. Kurgan, P. Musilek (2006); *A survey of Knowledge Discovery and Data Mining process models*. The Knowledge Engineering Review. Volume 21 Issue 1, March 2006, pp 1 - 24, Cambridge University Press, USA.

[Kushmerick2003] N. Kushmerick, B. Thomas, *Adaptive Information Extraction: Core technologies for Information agents*, Springer, s. 79–103, 2003.

[Madria1999] S. K. Madria, S. S. Bhowmick, W. K. Ng, F. P. Lim, *Research issues in Web data mining,* in: Proceedings of Data Warehousing and Knowledge Discovery, First International Conference. DaWaK'99, pp. 303-312, 1999.

[McGuinness2004] D. L. McGuinness, F. van Harmelen, *OWL web ontology language overview* (W3C recommendation 10 february 2004), 2004.

[Mulpuru2011] S. Mulpuru, V. Sehgal, P. F. Evans, D. Roberge, Forecast: US Online Retail Sales, 2010 to 2015, Forrester Research, Inc., 2010.

[Page1998] L. Page, S. Brin, R. Motwani, T. Winograd, *The PageRank citation ranking: Bringing order to the web*, Technical report, 1998.

[PIU2011] <http://poznajdirect.pl/direct-w-polsce.html>. Retrieved: 2017-06-01.

[PRNews2017] *Rynek bankowości mobilnej – IV kw. 2016*. <http://prnews.pl/raporty/raport-prnewspl-rynek-bankowosci-mobilnej-iv-kw-2016-6553798.html>. Retrieved: 2017-06-01.

[Quinlan1996] J. R. Quinlan, *Improved use of continuous attributes in c4.5*, Journal of Artificial Intelligence Research, 4, pp. 77-90, 1996.

[Ronka-Chmielowiec2003] W. Ronka-Chmielowiec, *Modelowanie Ryzyka w Ubezpieczeniach*, Wydawnictwo AE we Wrocławiu, 2003.

[Rumelhart1986] D. E. Rumelhart, , J. L. McClelland, & the PDP research group. *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press, 1986.

[Salam2003] R. Salam, *Estimating the Cost of Commercial Airlines Catastrophes A Stochastic Simulation Approach,* in The Casualty Actuarial Society Forum Winter 2003 Edition Including the Data Management Call Papers and Ratemaking Discussion Papers, pp. 379.

[Srivastava2000] J. Srivastava, R. Cooley, M. Deshpande, P. Tan. *Web usage mining: discovery and applications of usage patterns from web data*. ACM SIGKDD Explorations Newsletter, 1(2):12–23, 2000.

[Stolarski2015] P. Stolarski, *Metoda ekstrakcji modeli wyceny składki ubezpieczeniowej ze źródeł internetowych*. PhD thesis, UEP, 2015.

[Werner2010] G. Werner, C. Modlin, *Basic Ratemaking* 4th ed., Casualty Actuarial Society, 2010.