PwC Actuarial Services Newsletter

Issue 6
Advanced Analytics

Key points in brief:

- *Article #1: Data driven KYC: Know Your Customer and transaction monitoring for financial institutions*
- *Article #2: Machine learning and text mining: Enabling text data for use in models and software development*
- *Article #3: Smart Price Architecture: How insurers can modernize the structure of their market prices to enhance their profitability*

Advanced analytics - editorial note

The pace of change in the Financial Services industry is accelerating. New products and services are being developed and emerging technologies, such as Artificial Intelligence (AI) and Internet of Things (IoT) evolve the way of doing business.

According to our Global FinTech Report 2017, the increased sophistication of data models and analytics to better identify and quantify risk in the insurance industry is seen as the most important trend and the one to which the market is the most likely to respond. For this reason insurers are embracing innovation with a focus on advanced data analytics and 84% are planning to invest in it in 2017. We see a similar trend in the asset management and banking industry.

In this newsletter we show you three practical use cases in which we supported our clients to apply advanced analytics techniques to various business problems. They describe not only the algorithms and techniques used, but also the added value of the insights created and how to embed them into the business.

We hope you get inspired by these success stories. Please contact the authors of the articles in case you would like more information on how to bring these to your local markets.

pwc

Article #1: Data driven KYC: Know Your Customer and transaction monitoring for financial institutions

Supervisors are pushing for clear, up-to-date and accurate client monitoring, but are the financial institutions ready to update their business and harvest the associated wins?

Contacts

Marvin Oeben
marvin.oeben@pwc.com

Tom van der Vorst
vorst.tom.van.der@pwc.com

Introduction

This year, the Dutch Central Bank (DNB) is increasing its focus on the Product Approval and Review Process (PARP) for all Dutch financial institutions.[1] According to the supervisor, financial institutions should incorporate more data driven methods in their client and transaction monitoring in order to prevent the financing of terrorism, money-laundering and violations of the Sanctions Act.

Data driven techniques need to be implemented to show the supervisor that the financial institution is in control of these processes. These techniques include rule-based techniques from business experience and prescriptions from the supervisor. They should be enhanced with anomaly detection and network-based algorithms to ensure the institution can quickly identify new suspicious behavior and act accordingly.

The increase in availability of data, computing power and focus of the supervisor on these topics will lead to an increase in data driven methods for Know Your Customer (KYC) and transaction monitoring for financial institutions. In this article we lay out the required steps and their implications in the banking and insurance industry.

KYC, and why it's important

In the Netherlands, DNB assesses and enforces the adequacy and effectiveness of the procedures and measures implemented by supervised institutions to combat money laundering and terrorist financing as part of their integrity supervision[2].

For banks and insurers, it is essential to have an ex ante expectation of the profile and behavior of their customers. Based on that knowledge, they should determine which group of clients should be closely monitored and be able to identify when a client diverges from its expected (transaction) behavior. In short, the institution should:

1. *Know your customer* – know who (or which kind of company) your customers are, their profile, their actions and behavior and be able to predict their future actions.
2. *Know your customer's risk* - assign risk levels to customer profiles and routinely update these. Your PARP and customer approval process need to be up-to-date and one should actively monitor customer behavior. You should be prepared to find new, unknown risks even in small or old portfolios.
3. *Act according to the risks* - have scenarios, triggers and actions in place for determining actual undesired customer behavior and provide feedback throughout the business.

To assess the current effectiveness of the monitoring system in the Netherlands, DNB has performed research on this topic throughout the financial sector[3]. They have raised their concerns regarding the quality of the transaction monitoring process at financial institutions and their ability to observe and report suspicious transactions to the Financial Intelligence Unit (FIU). Moreover, they state that the risk profile and classification of clients is often not sufficiently taken into account in the scenario setting.

Although supervision is to a large extent coordinated by local supervisors, an effective supervision cannot be solely focused on the local market. Globalization and the improved economy have had an exponentially increasing effect on the number (and size) of the international transaction that are made.

[1] Toezicht Vooruitblik 2017 – De Nederlandsche Bank
[2] DNB Guidance on the Anti-Money Laundering and Counter-Terrorist Financing Act and the Sanctions Act
[3] DNB newsletter, August 2016

The increasing number and size of Dutch foreign transactions

(a)

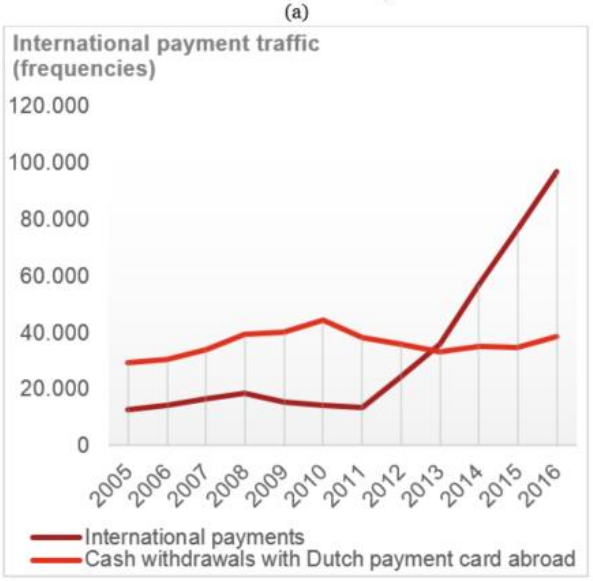


Figure (a): The significant increase in the number of international payments in the past years. Note that the steep increase already started before the introduction of the Single European Payment Area (SEPA) in 2014. The number of cash withdrawals has remained stable over the same period.
Source: DBN website – Retail betalingsverkeer.

(b)

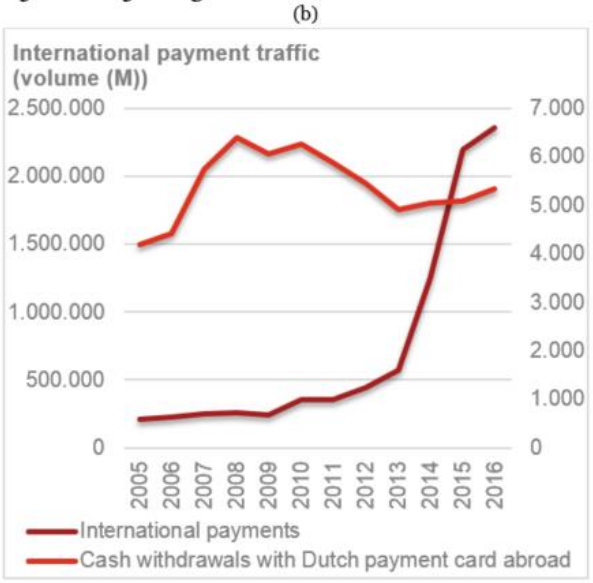


Figure (b): On the left axis the increase in the international transaction volume during the past years is displayed. The right axes shows the number of withdrawals. Moreover, the average transaction size has increased from €16 (2005) to €184 (2016). The amount of cash withdrawals have been more stable than the international payments.

Because of the international nature of payment traffic, the European Parliament has created regulation to set out the European vision on how transaction monitoring should be implemented. They acknowledge in the Money Laundering Directive that *"the changing nature of money laundering and terrorist financing threats, facilitated by a constant evolution of technology and of the means at the disposal of criminals required quick and continuous adaptations of the legal framework"*. These adaptations are needed according to the Directive in order to efficiently address existing risks and prevent new ones from arising.

But how could the monitoring system be improved without having to significantly increase the size of the teams involved?

Methods for monitoring

Although supervisors state that risk profiles and expectations with respect to the transaction behavior should be incorporated in the monitoring, there is no prescribed method to effectively do so. However, the guidelines do provide enough information on the expectations of the supervisor. For example, there should be a feedback loop in the monitoring and abnormal (and possibly suspicious) transactions should be observed and reported.

This would translate into a combination of the following techniques which together can create a more effective system:

1. Scenarios and/or business rules with profile specific thresholds.
2. Supervised learning to learn from previous (known) undesired behavior which will lead to tested and improved scenarios.
3. Unsupervised learning to combine (i.e. cluster) customers with similar behavior and characteristics and to detect abnormal customer behavior. The outcome of these techniques can then be incorporated in the scenario generation and testing process.
4. Network analysis can find cash flows, connected customers and central money transfer hubs which may be indications of abnormal behavior.

Besides these techniques, having up to date customer information such as the country of residence for private customers or the type of business branch for commercial clients should be considered basic KYC information.

Scenarios with profile specific thresholds

A situation in which there is risk that a client is performing unusual or illegal activity is translated into a defined scenario. Scenarios make use of threshold settings, which determine when transaction behavior is considered to be either suspicious or should be reported to the supervisor straight away. Thresholds within a scenario can be set at three levels of complexity:

1. Low complexity – Low complexity thresholds are based on a single variable, or a combination of single thresholds for multiple variables (horizontal or vertical lines).
2. Medium complexity – A more sophisticated method of setting thresholds to allow for a combination of multiple variables (straight lines that could be a linear combination of more variables).
3. High complexity – Advanced thresholds could be implemented as a union of convex hulls, which would result in a scenario setting that is closely related to the expected behavior of the customer.

Increasing the complexity of the threshold setting and the accuracy of the transaction profiles to which the scenarios are applied can lead to a large reduction of false positives. Note that using a more sophisticated method can also lead to challenges in the implementation of these custom scenarios in transaction monitoring systems.

Different levels of complexity in threshold setting

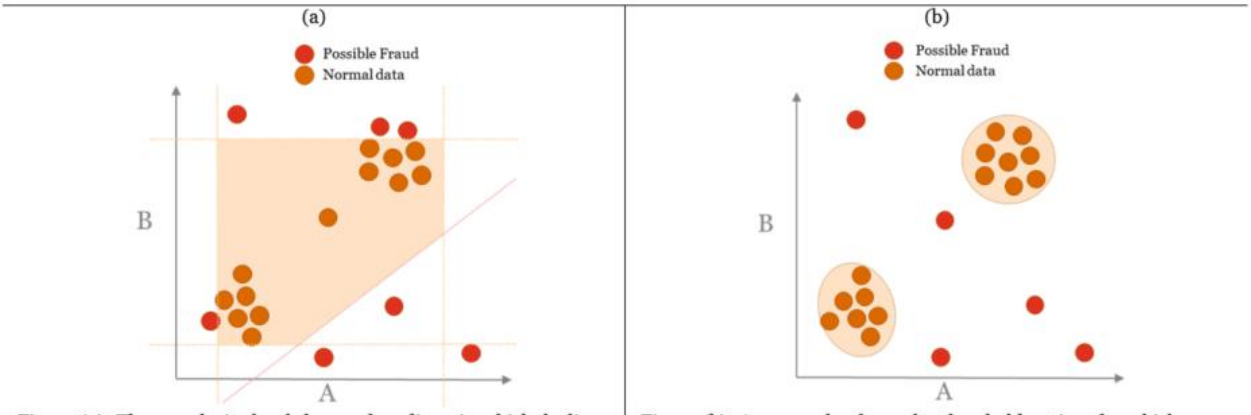


Figure (a): The complexity levels low and medium, in which the lines indicate the thresholds. For a profile specific implementation the straight lines can be set per segment.

Figure (b): An example of complex threshold setting, for which per segment a convex hull is defined that is closely related to the actual behavior of the clients associated with the profile.

Effectively detection of unusual behavior can be done separately for different customer profiles. A credible number of customers inside a single profile (based on historical data) can then insure that proper thresholds can be found. Thresholds can be set using the quantiles of the observed distribution. Note that in case of a scenario that is based on multiple variables this can also be applied to the multivariate distribution.

Supervised learning to learn from previous (undesired) behavior

Knowing which client behavior has led to identified fraud in the past is valuable information. Considering the demand for feedback required by the supervisor, this information can be used to score and predict fraud in the future. Especially in an anti-money laundering (AML) setting this can significantly improve the current systems in use. Moreover, by using fraud confirmed by the FIU one can achieve a large reduction in false positives while not losing a significant number of true positives.

Furthermore, one can analyze the alerts that were generated by scenarios that never resulted in an FIU confirmed fraud and improve the threshold settings such that the absolute number of alerts reduces as well. In this way, supervised learning will reduce the absolute number of alerts as well as the (relative) number of false positives.

Unsupervised learning to find similar or abnormal behavior

From a KYC perspective, customer profiles should be based on both characteristics and behavior. These profiles can be created using unsupervised (clustering) techniques. One could use a similar approach in AML detection, but with different characteristics.

Once clients with similar behavior belong to one profile, it is possible to set better (i.e. more accurate) thresholds on possible suspicious behavior. The combination of clustering and thresholds can be easily implemented in the current client's systems, as only the inflowing groups need to be redefined and thresholds recalibrated.

A supporting technique to clustering is anomaly detection, which detects abnormal behavior within the data. The concept is such that when the normal behavior within the data is known, we can use a metric (e.g. distance based, probabilistic) to determine abnormal transactions or client behavior. We are, however, not restricted to our predefined clusters but may run anomaly detection on its own to detect abnormal behavior without a predefined bias.

Network analysis to find identify suspicious money flows

Recent technologies have empowered researchers to find network structures which may indicate suspicious behavior. Examples of this are regular transfers between seemingly unrelated market segments or so-called 'sinks', which are points in the money flow which money flows into but then disappears (in cash or to an outside bank). If these sinks are deemed suspicious then so should their funders. This analysis can be used both in anomaly detection as well as the supervised feedback loop.

Furthermore, a network analysis on transactions can be used to identify relationships between two customers from different channels or a social community. An article written by Molloy I. et al (2017)[4] even shows an application in which a network analysis is successfully used to identify fraudulent transactions and help in cross channel fraud detections.

Look into the future.

DNB has stated that terrorist financing, money-laundering and the sanction act will be part of the supervisory agenda. Enhancement of the KYC and PTM process should therefore be on the top of the agenda of all financial institutions. The time is now to move the processes forward and embrace machine learning and artificial intelligence as a powerful tool to streamline their business. Using a combination of techniques, the KYC and PTM process can be improved without increasing the size of the teams that are responsible for those processes.

[4] Molloy I. et al. (2017) Graph Analytics for Real-Time Scoring of Cross-Channel Transactional Fraud. In: Grossklags J., Preneel B. (eds) Financial Cryptography and Data Security. FC 2016. Lecture Notes in Computer Science, vol 9603. Springer, Berlin, Heidelberg

Article #2: Machine learning and text mining: Enabling text data for use in models and software development

Contacts

Herbert Julius Garonfolo
hjg@pwc.dk

Simon Kirkeby Wessel
skw@pwc.dk

Introduction

As data analysts, we love to build our models on nicely structured data. However, we often see our clients store important data as unstructured text in reports, emails, logs, text boxes and comment fields without any validation or structure.

In this article, we will describe how to transform large amounts of unstructured text, as found across CRM and ERP systems, into structured text that can be used in models and software development. We will do this using a supervised machine learning model that can handle spelling errors and other noise in text data.

Possible Business Cases

As a business case, imagine an insurance company that receives massive amounts of text in the form of customer health records from healthcare providers. The insurance company is trying to understand these health records and classify them as high, medium or low risk, or alternatively find out which health issues are mentioned in the records.

It could also be that an insurance company receives damage reports from a car workshop and needs to understand which types of damage and necessary repairs are described in the damage reports.

Normally, in both examples it would require significant human resources to go through all the data manually and classify the text snippets, one by one. Additionally, the text cannot be used in statistical models or be treated logically by software applications as long as it is unclassified and unstructured.

KNN Example Using Twitter Data

Luckily, this kind of problem can be solved using machine learning to classify the text. In the following section, we will give you an example of a supervised machine learning K-nearest neighbor (KNN) text classification model that we use to classify text snippets from two different Twitter accounts.

The first account is @EBA_News which is the Twitter account of The European Banking Authority. EBA is the EU agency that works to safeguard the integrity, efficiency and orderly functioning of the EU banking sector.

The second account is @actuarialpost which is the Twitter account of the online actuarial magazine The Actuarial Post.

Text data is often more complicated to analyze compared to numerical data. However, unstructured text input might contain very valuable information and advanced data analytics is a powerful tool to extract this information. This section presents the KNN algorithm, which we use to classify text.

In this example, we are trying to determine which Twitter account a particular tweet came from. In other words, we have a dataset of tweets split into two parts: a training set, where we know which account the tweet came from, and a test set where the account is unknown. We will use the KNN model to try to determine which account the unknown tweets came from.

KNN is a supervised machine learning method, used for classification. A KNN model predicts the target output of new un-categorized observations. New observations are compared to a number, K, of similarly categorized observations.

Before the tweets can be compared a “corpus” of words is defined. The corpus contains all the significant words from the tweets, i.e. text that has been standardized and cleaned for common words. The text clean-up is done by removing “stop words”, e.g. me, on, with etc., and special characters like “@” and “#”.

The next step is to identify the stem of words. This means that you remove all the endings resulting from abbreviation and subject-verb agreement (i.e. singular or plural form of the word). As an example, instead of treating “insured”, “insurance” and “insurances” as three different words, we will stem all these words to “insur” and treat them as one word. By performing this on each word, we end up with a document term matrix (DTM) of word stems from the corpus. This improves the accuracy of the model.

The DTM is generated by counting the number of times that each word in the corpus occurs in each cleaned tweet ("the documents"). Each row contains the frequency of words that occur in a certain document. Rows in the DTM are then compared to calculate the similarity of tweets. This allows us to compute the Euclidian distance between all tweets, which in turn allows us to identify neighbor tweets (i.e. similar tweets).

For the example, in this article, we used Python to download and create a small dataset consisting of 600 freely available tweets from EBA and The Actuarial Post; and R to create and run the KNN model.

The corpus created from the tweets consists of 1,944 cleaned words occurring 6,748 times. Plot 1 shows the most common terms found in the tweets from both accounts. "New" is the only term used frequently on both accounts. Unsurprisingly, the most common terms used on the separate accounts relate to their industry, i.e. "pension" and "bank". Categories using the same terms more frequently, are harder to separate. Therefore, removing the term "new" from the data, might improve the performance of the model.

A test data set is defined using 20% of the tweets from each category. The KKN model is trained using the train set and then used to categorize the tweets in the test set. If we configure the model to use 10 neighbors in the classification, the model predicts the source of a tweet with an accuracy of 87%.

The confusion matrix shows how the 600 tweets were classified, as well as the performance of the classifications. More tweets are predicted to come from the Actuarial Post but 20% of these predictions are false. The accuracy of the model is significantly higher when a tweet is classified as an EBA tweet.

	Actual	
Predictions	Actuarial post	EBA
Actuarial Post	300	74
EBA	2	224

By varying the number of neighbors, the performance of the model might be improved. The accuracy of the model using a varying number of neighbors is shown in plot 2. The performance is lower when a smaller K is used. The best performance is 93% and found for K=13.

The sparsity of the DTM is almost 100%, meaning that many words appear only in very few of the documents. Removing terms occurring rarely in the DTM can reduce sparsity and might lead to a better accuracy. However, the train set had only 150 tweets, so in our case, we would be better off by increasing the sample size. Using 50% of the tweets for training showed significant improvement to accuracy.

When we applied a KNN model to determine the source of a tweet we obtained an accuracy of 93% using 13 neighbors meaning that the model categorized 93% of the test data correctly.

We can conclude that by training a KNN model, it is possible to classify and categorize massive amounts of text data. These categories can further be used as variables in statistical models and for other software development purposes like database logic etc.

The Future

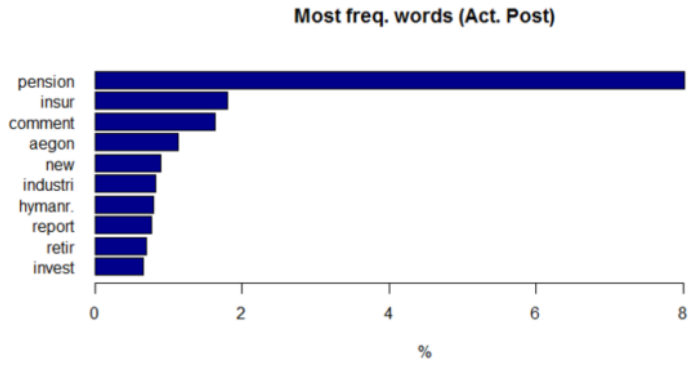
As we mentioned in the beginning of the article, a simple method like KNN has a number of interesting business applications. If we return to the example of the insurance company receiving health records, imagine how this company could significantly reduce the resources used on reading, understanding and categorizing these health records by using a KNN model to classify and categorize the records.

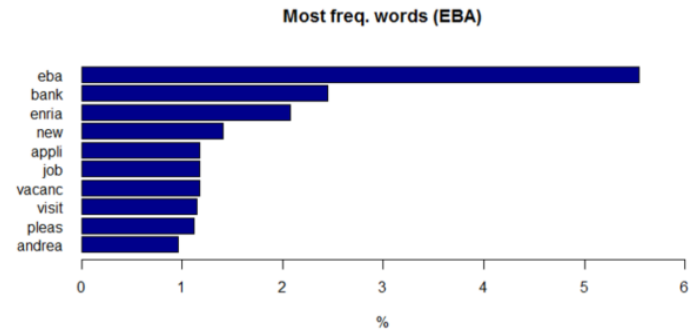
We do not propose to completely remove humans from the process but in many cases, a KNN model can considerably reduce the resource strain by pre-treating the data, thus allowing humans to be much more efficient.

Furthermore, the structured data coming out of a KNN model is much easier to use in the traditional GLM models. Therefore, using a KNN model as a preliminary model for the traditional GLM, one is able to use almost any kind of text data as a variable.

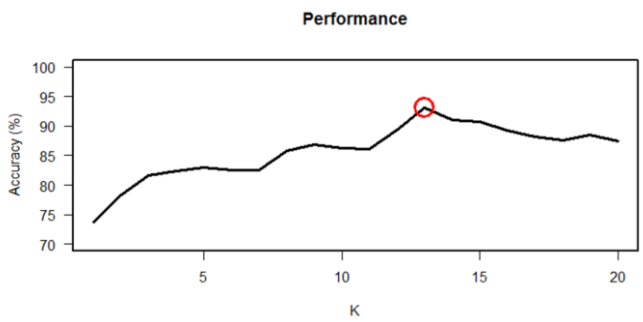
If you have any questions or comments, please feel free to contact us.

Plot 1





Plot 2



Article #3: Smart Price Architecture: How insurers can modernize the structure of their market prices to enhance their profitability

Contacts

Frank Schönfelder
frank.schoenfelder@de.pwc.com

Alexandre Veber
alexandre.veber@fr.pwc.com

Maximilian Hudlberger
maximilian.hudlberger@de.pwc.com

Introduction

These days, digitalization and the use of Big Data has caused a lot of changes in many parts of the insurance industry. Pricing is on the forefront of where this can be observed. One might wonder how data science techniques could revolutionize pure premium models and render obsolete the Generalized Linear Models (GLMs) which are widely used across non-life products. To take this thought experiment a step further one could also wonder how to enrich these models with competitors' pricing data to impact price competitiveness.

The PwC Smart Price Architecture combines classical GLM and modern machine learning techniques. It uses new risk models, reverse engineering of competitors' prices and stacking algorithms.

The following gives a short overview of this architecture illustrated over a public data challenge launched in July 2016. By implementing this approach we achieved great results in terms of segmentation capability. This gives a significant competitive advantage by better managing the anti-selection phenomenon.

The Actuarial Pricing Game

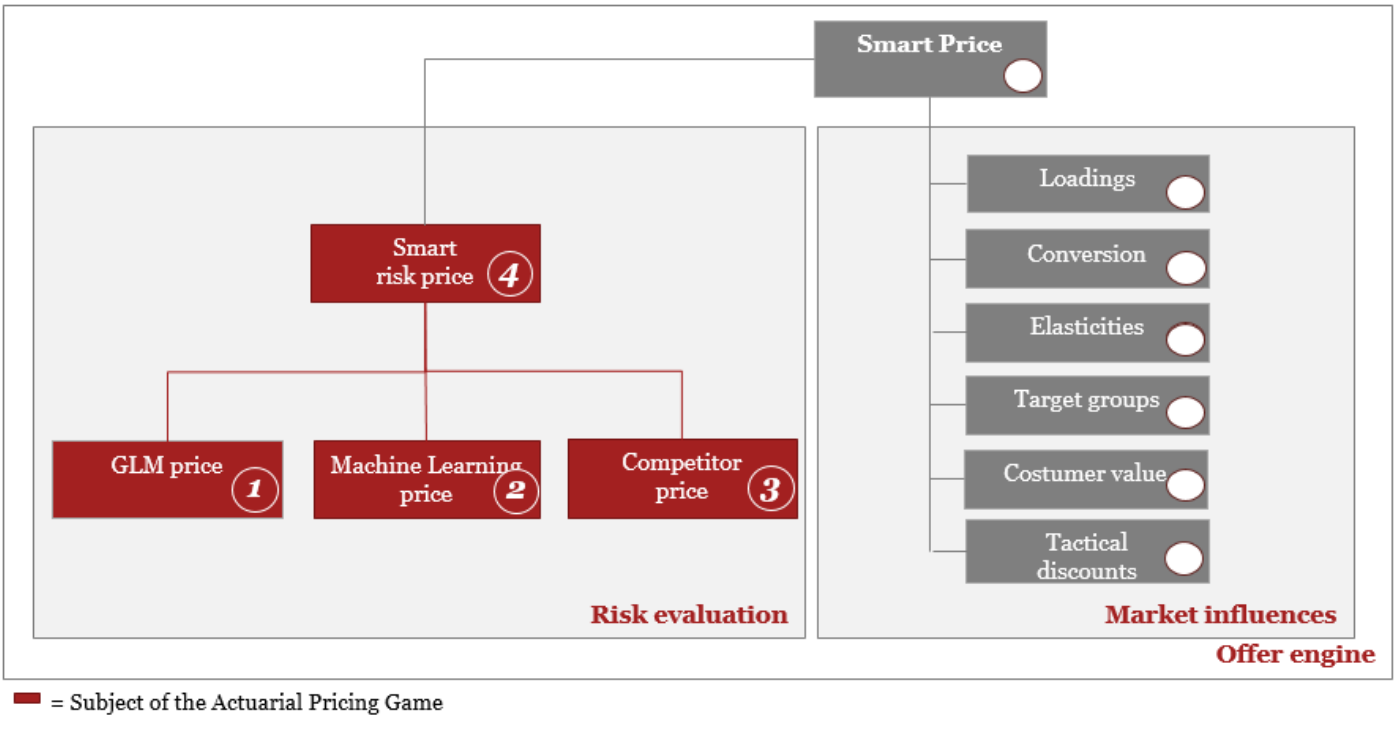
The second Actuarial Pricing Game, organized by Arthur Charpentier and the French society of actuaries was launched on July 15th 2016. A motor insurance data set, containing the years 2011 – 2014 was provided. The aim was to provide fair premiums for the test set (2014).

The Pricing Game was split in two phases. In phase one every player had to deliver a first set of premiums for the test set. For the second phase, each player was given a part of the premium of two competitors. Everyone was allowed to adjust their premiums by taking the competitors' ones into consideration.

The purpose for this exercise was to arrive at a better risk selection approach when compared to competitors.

Methods used

Our strategy was to implement the Smart Price Architecture, which mainly consisted of four steps.



GLM price

1

In the first phase we built a **GLM** using Penalized Regression. GLMs are easy to implement and penalized regression help in aiding modeling efficiency. The method we used was an elastic net. It is a combination of LASSO and Ridge penalization methods. There are two main advantages of this method. On the one hand, it can be used as an effective and fast feature selection tool for a GLM. On the other hand, the penalization helps to prevent overfitting. With this model we were able to give a suitable risk prediction for phase 1 which nicely handles real life operational constraints.

Machine Learning price

2

In the second phase we built additional models, e.g. a *gradient boosting machine (GBM)*. One great advantage of these **machine learning models** is the efficient handling of dependencies and the automatic detection of patterns. Furthermore, they are better at handling high dimensional data. We used these models as sub models for our final prediction in our last step.

Competitor price

3

One complex piece of work was to use the partial information given about competitor premiums. As we only received partial information, we had to rebuild their pricing algorithms. To handle this task we used **reverse engineering** techniques. The method we implemented here was again the *GBM* as this is a commonly used technique in such data challenges. In addition, it has some nice properties which were beneficial for the task.

We were not predicting risks, but closed formulas, as we were predicting the competitor models. For this reason, we over fitted our model by creating a large amount of trees. In general this is not recommended as it learns the training data too closely. In our case however, there should be nearly no randomness in the data, so it can simply learn the model formula. As a result, we obtained **nearly exact estimates of the competitor premiums**. Just as the risk models in step one, they represented sub models for our final model afterwards.

Smart risk price

4

By the end of phase two, we had created our own multiple risk models (like the LASSO and GBM) as well as two competitor models. Each of these models could be used to give a risk prediction for the test set. Even if one of them was technically the best performer on a given test set, it won't necessarily be the best in every part of the test set. This is even more true in real life where each competitor owns only a sub part of the industry wide insured data base.

This is where the last part of our **Smart Price Architecture** came into play. We **combined the predictions** in such a way that we took the best out of each sub-model. We achieved this by using stacking techniques.

Stacking is an intelligent way to combine different models. The method itself could, for example, be just a weighted linear model, a GLM or a more complex model such as a GBM or a neural network. If using a GLM the stacking technique just uses the different sub models as predictors in the design matrix. The consequence is that each model has an influence on the overall prediction with different impact given their ability to predict sub parts of the data set. In reality, the Smart Price Architecture can be fed by even more models like customer value models or price elasticity models.

Results

By using the different models and aggregating them with stacking we not only got our best model but the **best overall model of all competitors**[1] regarding the gains curve and the related Gini Coefficient. These indications stress the ability for models to classify risk by order of magnitude. The closer to 1 the better the Gini.

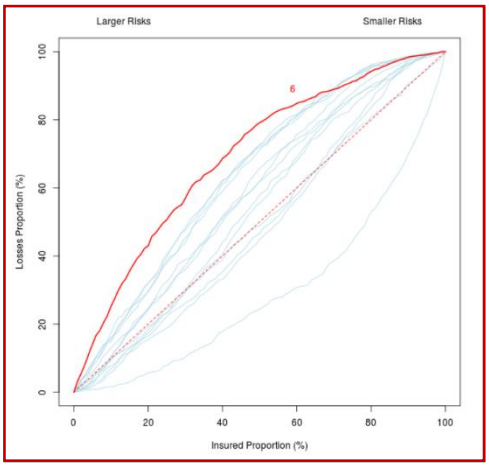
What was also interesting is the fact that competitors played the game to make reductions which were too expensive and increases which too cheap. The result for them was a drastic drop into their risk segmentation ability exposing themselves to a large anti-selection problem. In comparison, the **Smart Price Architecture efficiently reduces anti-selection.**

Conclusion

The results of the Pricing Game as well as the practical experience with our clients show that the Smart Price Architecture is a powerful way to enhance risk segmentation. With this architecture, insurers can use the important new techniques available on the market, preventing anti-selection and staying competitive.

The insurer doesn't have to discard any currently used models but instead use them as an input for the stacking model.

When implementing the Smart Price Architecture, insurers have to consider an efficient way of deploying the price in their marketing efforts.



[1] *Source: https://de.slideshare.net/charthur/pricing-game-100-data-sciences, Slide 15, Insurer 6 = PwC*

Actuarial Services – Your Contacts

The Netherlands

Jan-Huug Lobregt
Tel: + 31 88 792-7463
jan-huug.lobregt@pwc.com

Bas van de Pas
Tel: +31 88 792-6989
bas.van.de.pas@pwc.com

Theo Berg
Tel: +31 88 792-2623
theo.berg@pwc.com

Marc Bout
Tel: +31 88792-6068
marc.bout@pwc.com

Denmark

Jette Lunding Sandqvist
Tel: +45 3945-3817
JLS@pwc.dk

Germany

Clemens Frey
Tel: +49 895 790- 6236
clemens.frey@pwc.com

France

Emmanuel Dubreuil
Tel. : +33 1 56 57-19 37
emmanuel.dubreuil@fr.pwc.com