Towards a data-driven approach for fraud detection in the social insurance field: A case study in Upper Austria

Johannes Himmelbauer[1], Jorge Martinez-Gil[1], Michael Ksen[2], Katharina Linner[2], Sieglinde Plakolm[2]

[1]Software Competence Center Hagenberg GmbH
Softwarepark 21, 4232 Hagenberg, Austria
-
[2]Oberösterreichische Gebietskrankenkasse
Gruberstrasse 77, 4020 Linz, Austria
-
e-mail: Johannes.Himmelbauer@scch.at

Abstract. The Social Insurance industry can be considered as a basic pillar of the welfare state in many countries around the world. However, there is not much public research work on how to prevent social fraud. And the few published works are oriented towards detecting fraud on the side of the employees or providers. In this work, our aim is to describe our experience when designing and implementing a data-driven approach for fraud detection but in relation to employers not meeting their obligations. In fact, we present here a case study in Upper Austria but from which interesting lessons can be drawn to be applied in a wide range of different situations.

Keywords: Data mining; health insurance; social insurance; fraud detection;

1 Introduction

In many countries of the world, social insurance plays a crucial role according to the social security and welfare of a state. The Austrian health system is based on the principles of solidarity, affordability and universality. Many of its responsibilities have been delegated to self-governing bodies, although most of the regulations are defined by law. Therefore the Austrian health system is complex, fragmented, and because of its high standards it is relatively costly [3]. The Austrian health insurance is mainly financed by income-related periodic contributions the employers have to pay.

This system has proven to work very well and high levels of well-being and safety for workers have been achieved. However, the rigorous supervision of the whole system is necessary so that degradation does not occur. In this context, fraud control is of great importance as a means of keeping down the costs for employers who fulfill their obligations properly. The major problem here is that

fraud is intended to be processed as normal, which means that fraud has to be looked for to be discovered, and this task is far from being trivial. Moreover, most fraud control activities are often performed manually, this means that unexplored automatic fraud detection techniques have an enormous potential to make an impact on this context. For these reasons, this research intends to shed light on the design of a data-driven strategy to help to detect fraudulent behaviors. Therefore, the major contributions of this work can be summarized as follows:

- We introduce our data-driven approaches for fraud detection in the context of employers that do not follow the rules for hiring workers.
- We present a use case in the context of the Upper Austrian health insurance, whereby our approaches have shown to provide useful support to control fraudulent activities.

The remainder of this work is organized as follows: Section 2 outlines the State-of-the-Art in relation to fraud detection for Social Insurance industry. Section 3 introduces a real-world case study for fraud detection in the area of employment settled at the Austrian social insurance system. Section 4 presents the data-driven procedure that we have designed and implemented in order to facilitate and support fraud monitoring. Section 5 summarizes the achieved results with respect to the presented case study and how our approach have had an impact on the regional health insurances in Austria. Finally, we highlight the conclusions and future research lines that could be derived from this work.

2 State-of-the-art

There is a large body of literature in relation to fraud detection in the context of the Social Insurance industry, but in the vast majority of cases it is either in relation to workers who use dishonest tactics to get benefits that are not theirs, or it deals with service providers (e.g. doctors, hospitals) that make false claims by charging the insurance for unnecessary or even not done treatments. In general, it is widely assumed that the Social Insurance industry consists of the following stakeholders:

- Insurance carriers (governmental health departments, private insurance companies)
- Service providers (hospitals, doctors, laboratories,...)
- Insurance subscribers (employees and patients)
- Insurance payers (employers)
- External providers (pharma industry)

With this vast amount of actors in mind, it is not difficult to envision that there are multiple variants to commit fraud. In fact, Thornton et al. [11] have identified many different kinds of fraud, among which the following stand out: *identity theft*, i.e. stealing confidential information from stakeholders and using

that information to prepare false bills, *phantom billing*, i.e. billing for good or services that are not actually performed, *unbundling*, i.e. billing different phases of a procedure as if it was a different treatment, *upcoding*, i.e. billing services more expensives than the ones performed, *bill padding*, i.e. providing unnecessary services to a patient, *kickbacks*, i.e. a negotiated bribery in which some money is paid to do something in return, and many more.

In order to fight against frauds of this kind, some remarkable techniques have been already proposed. For example, Rawte et al. built a novel hybrid approach making use of supervised and unsupervised learning for detecting fraudulent claims [8], Diaz-Granados et al. have proposed a solution to extract and analyze social media data in pursuit of identifying insurance fraud [1], or Dua and Bais have worked towards novel data mining fraud detection models [2]. Slightly different is the proposal of Tsai et al., who have proposed a knowledge model along with the existing database applications using the popular CommonKADS methodology [12].

However, the case that we address here is of different nature, since we focus in companies that do not register (some of) their employees appropriately in order to reduce the labor costs, so we mean a situation shared by insurance payers, insurance subscribers, and insurance carriers. Therefore, this fraud directly harms both the employee (insurance subscriber), who cannot make use of his or her right to appropriate medical treatment and the insurance company (insurance carrier) that does not receive the proper contributions that all employers are obligated to pay for having people working for them.

When analyzing the literature in this context, it is possible to see that most of the works belong to one of these two large groups: those that focus on issues such as causes, consequences, statistics, impact on society, etc. and those that describe techniques that can be useful to help to detect fraudulent cases. In addition, recent breakthroughs in computational paradigms such as artificial intelligence, data mining, and machine learning allow many of these techniques to be implemented (and even improved) by means of computer systems. In our particular case, we are interested in describing our experience in relation to the research and development of some of these fraud detection techniques by means of computer systems.

In relation to the existing literature in this field, interesting works have been carried by Van Vlasselaer et al. whereby the goal is to identify those companies that intentionally go bankrupt in to avoid paying their contributions [13]. Widder et al. proposed a fraud detection by using a combination of discriminant analysis and techniques based on artificial neural networks [14]. To do that, they propose an event processing engine for detecting known patterns and aggregating them as complex events at a higher level of analysis in real-time. Finally, Konijn and Kowalczyk propose a method that consists of analyzing the historical records and aggregating these results in order to detect outliers [4]. The work that we present here is, to the best of our knowledge, the first attempt to build up data models that can support the generation of recommendation lists based on current data so that fraudulent behavior can be inspected according to a ranking of priorities.

3 A case study - Social fraud detection in the area of employment

The Austrian social insurance system is based on the principles of solidarity, affordability, and universality. It is primarily funded through insurance contributions. It includes the branches of accident, health and pension insurance, and it is formed by a number of institutions existing under the Main Association of Austrian Social Security (HVB) as their umbrella organization.

In this context, the Austrian social insurance system constantly has to face the most diverse amount of attempted frauds. One major type of fraud is related to employers not meeting their obligations, i.e. they do not pay the proper amount of contributions for their employees. More concretely, either they do not register worker(s) at all (classical black labor), or - more often and also more difficult to discover - they specify a wrong, too low assessment basis for their payments.

Consequently, there is continuous work on measures against social fraud in that area. Over the years a steadily growing base of knowledge and experience has been build up by the financial in-house experts at the Austrian health insurance. This available expert knowledge is commonly used as follows to detect or even prevent social fraud in the area of employment: in-house experts manually examine the available data of selected companies and if according to the expert there are found reasonable suspicious circumstances in the data of a certain company it might ask the proper authority for inspections at site. For this purpose, the expert usually either searches for already known specific suspected patterns (e.g. in the construction sector a high rate of marginal part-time employees is suspicious) or checks if in general there can be found larger deviations from average behavior.

In the course of this process, during last years a dashboard tool has been developed at the Upper Austrian Health Insurance[1] whose aim is to support in-house experts in their work towards social fraud detection. More concretely, the RAD-Tool (German abbr. for risk conspicuousness of employers) enables the user to visualize and compare relevant historical and current data of each employer. Figure 1 contains a screenshot of the tool. Utilizing this dashboard in daily work facilitates and improves the examination of companies based on their data. However, one major issue remains: considering the available personal resources (usually only a few in-house experts per federal region) it is impossible to check a larger part of existing employers (up to 100.000 per region). Therefore selecting randomly companies for examination remains somehow like looking for the needle in the haystack. This is where our work intends to tie in and great potential with respect to automatic data analysis is seen. Basically, the motivation for the data-driven strategy that is presented in the following section is to provide the dashboard user with automatic recommendations (based on current and historical data) of which companies are considered worthy to give a closer look.

[1] Oberösterreichische Gebietskrankenkasse (short OÖGKK)

Fig. 1: Screenshot of the dashboard tool RAD (German abbr. for risk conspicuousness of employers)

We have chosen a semi-automatic, multi-step approach for building up models that automatically generate recommendation lists based on current data. *Semi-automatic* means that besides data-driven steps manual interventions (like final model selection) are necessary. The reasons why we have refrained from trying to install a fully automatic approach are manifold:

Use of prior knowledge: We want to incorporate the extensive expert knowledge that is available.

Legal conditions: The mere suspicion of fraud is not enough to act as legal proof, so a computer system cannot determine by itself what is a fraud or what is not.

Soft requirements for model valuation: We need to take into account that the model is not intended to be used as a black box, but serves as a decision making support for human experts. A necessary condition that the system can be helpful is that the user finds basic confidence in its decisions. Therefore interpretable, comprehensible models that at least partially represent the existing knowledge and intention of the user are preferable

to non-transparent, complex systems; even if the later show slightly better performance with respect to statistical performance measures.

In the following, we describe a procedure that takes into account the above conditions by combining heuristic strategies and statistical evaluations.

4 Procedure for Knowledge Discovery in the Area of Social Fraud Detection

A starting point of our work was data regarding over 60,000 companies in Upper Austria and that is accessible for the Upper Austrian health insurance. The database consists of more than 200 entries per company and month (columns) ranging from basic information about each company (e.g. location, economic sector) to financial information (payments, payment default, financial problems like past insolvency) and development (company size, monthly fluctuations) and structure (e.g. sex, age, mode of employment) of the employer's staff. Additionally, the experts of OÖGKK have compiled (and are continuously maintaining) a list of "suspicious" companies at which social fraud has already been detected in history. This list of about 750 firms helps the in-house experts to focus on companies that - based on historical experience - are expected to present a higher risk level with respect to social fraud than the average.

To summarize, from the data analyst's perspective, the data available for our case study consists in a multivariate time series (usually we considered a history of 24 to 30 months) for each of the over 60,000 companies. About 750 of these (i.e. 1.2%) contain the flag *suspicious*, the rest remains unlabeled. The basic idea of our data-driven approach is as follows: we aim to build up a model that gives, as a result, a recommendation list which contains at the top the - according to the data-driven model - most suspicious company, followed by the second most suspicious company, and so on.

The model basically should fulfill the following two conditions: The majority of the companies that are ranked in the top part of the scoring list should be firms labeled as *suspicious*. Calculating the proportion of already *suspicious* employers for the top part of the scoring list gives the possibility to measure and compare the models' performance from a statistical point of view. But beyond, the decisions of the model should be easily comprehensible and need to represent a good, useful basis for final judgment by the expert. Amongst other things, this requires that the in-house experts get recommended companies from unlabeled data (i.e. still *unsuspicious* companies) that turn out to be interesting to be given a closer look. The following three subsections describe our way towards such a model.

4.1 Feature Generation

When starting this work in-house experts have already been using over years the available data to manually search for suspicious patterns indicating possible

fraud. For this, they do not look exclusively on data at single points of time, but they investigate the behavior and development of figures over a long past time period (usually one up to two years). There are basically two different approaches to build up a data-driven model in such a setting. Either the raw multivariate time series serve as input itself (e.g. cluster algorithms based on measuring the distances between time series), or aggregated features calculated from the time-series are used as input, instead. In the latter case, extracting potentially useful features prior to the training of a model usually turns out to be a crucial task. For our case study, we tried to make use of existing expert knowledge for doing so. Based on the documentation, explanations, and examples from the side of the experts we tried to come up with mathematical or rather statistical formulations to describe known suspicious patterns. In this way, a large set of potentially interesting features (several hundred) could be generated from the time-series data.

In the following, we show with a concrete example how the feature generation process was typically conducted: In-house experts can say from their experience that an unusually high fluctuation of staff over a longer period potentially indicates a committed fraud. Therefore, when examining a company, amongst other things, experts usually give a closer look at the development of the companies' staff. Figure 2 shows plots of the - in this regard - relevant time series for two different companies. On the left side, for *company A*, the monthly company size (red solid line) remains fairly stable over the whole observation period from January 2017 to December 2018 (with the maximum value of 20 employees). Only in 9 out of 24 months minor changes of the staff (i.e. registrations (turquoise dash-dotted line) and/or deregistrations (green dashed line)) were conducted. There is no odd behavior with respect to the staff's development observable. On the contrary, the plot on the right-hand exhibits major changes. Considering only the time series of the company size (red solid line) with its rapid increase during the first observation year and a relatively stable number of employees during 2018 the behavior of *company B* could be explained as typical for a dynamically growing enterprise. Even the temporary reduction of employees during winter 2017/2018 is comprehensible due to the seasonality of the construction sector. However, a more detailed look to the monthly registrations and deregistrations statistics reveal that over the whole observation period a high proportion of the staff is continuously exchanged. For example, there is only a minor increase of the size of *company B* from 21 to 22 employees in November 2018, but, actually, a high number of change requests, namely 13 (7 registrations and 6 deregistrations), are recorded at the side of the social insurance. According to the experts' opinion, such behavior is abnormal and suspicious.

In summary, according to the experts, the available monthly data about the development of the staff (as shown in Figure 2) contains useful information to estimate a company's fraud risk. The example above suggests that rather than simply considering the development of the company size itself it is preferable to use the monthly numbers of both the registrations and deregistrations. A first statistical check whether the available data really confirms the above assumption

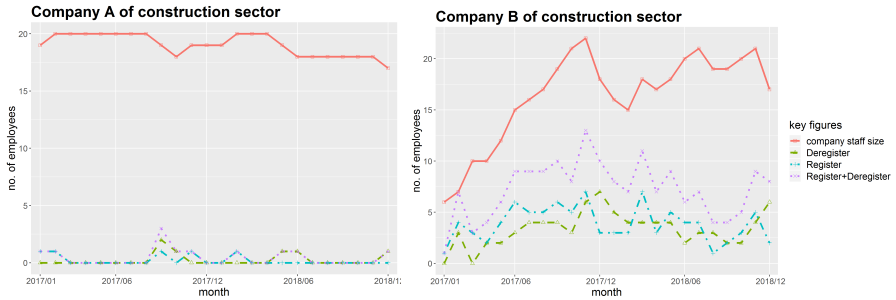

Fig. 2: Development of staff for two companies of the economic sector of construction (January 17 - December 18)

is to compare the average behavior of the already suspected cases with the - up to now - unsuspected ones. In order to obtain more comparable numbers among companies of all sizes it was decided to not use the absolute numbers of the monthly registrations and deregistrations, but to normalize them with respect to the current company size. Therefore, for each company c and month m the current fluctuation of staff is calculated by

$$Fluct(c,m) = \frac{nReg(c,m) + nDereg(c,m)}{company\_size(c,m)}, \tag{1}$$

with $nReg(c,m)$ and $nDereg(c,m)$ as the number of registrations and deregistrations at company c in month m and $company\_size(c,m)$ as the number of currently registered employees, accordingly. Figure 3 shows the monthly fluctuation averaged over all already suspected employers (red thick dash-dotted line) as well as averaged over all the other (unlabeled) companies (blue thick solid line). Hence, the blue line so to speak represents the normal behavior of a company within the construction sector.

Throughout the whole observation period, the average fluctuation of the *suspicious* cases is significantly higher than for the unlabeled companies (about double as high!). This means that the statistical evaluation strongly substantiates the existing experts' view described above. However, we want to point out that due to the high distributional variance of both groups (suspected versus still unsuspected companies) a distinct classification solely based on the staff's fluctuation (1) will not be possible (see the highly overlapping standard deviations shown in Figure 3). In other words there exist also other reasons than fraud why a company might exhibit a high fluctuation rate of employees, and vice versa.

In order to use the information that is contained in the time series of Figure 3 with respect to fraud risk for the model generation process we need to calculate aggregated features from the time series data. To sum up, in this situation we are in search for indicators that have high values in the case that a company's data exhibit a suspicious behavior like e.g. *company B* in Figure 2, and a low value

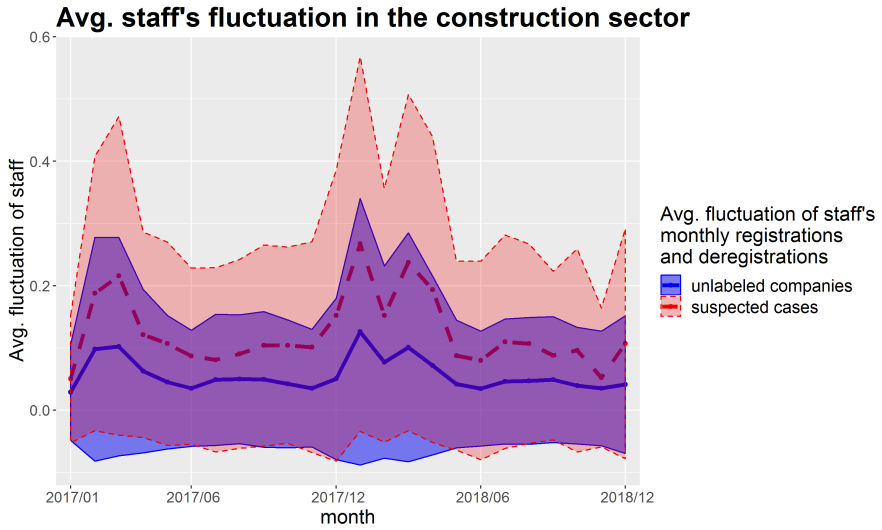

Fig. 3: Comparison of the average fluctuation within the economic sector of construction (January 17 - December 18): suspicious companies (red dash-dotted) vs. unlabeled companies (blue solid)

otherwise. Obvious feature candidates for such indicators are standard statistics (like average or extremal values) extracted from the corresponding time series data.

In practice, during our case study the feature generation process was typically conducted in the following way: First, we try to deduce a proper time series (by transformation and/or combination of time series available in the raw data) that lets reveal best a certain suspicious pattern described by the expert (like the staff's fluctuation (1) for the example described above). Next, we extract a set of standard statistics (typically *mean*, *median*, *max*, *min*, and *stddev* for numeric time series, and existence and number of occurrences for binary variables, respectively) over the whole observation period. In some cases, when considered meaningful, we additionally focused also on sub-periods like for example the winter season in case of the gastronomic sector in regions with ski areas. Such decisions were always resulted of taking into account both the expert's view as well as statistical indications.

4.2 Feature Selection

The result of the feature generation process described in the preceding subsection was a diverse set of possible input candidates for generating a model that estimates the risk level for each company. Instead of passing all these several hundred features directly to a learning algorithm it was decided to conduct the first preselection with respect to statistical significance and (visual) interpretability

of the features. This was motivated by our aim to finally achieve the highest possible comprehensibility for our model.

As already mentioned above, we tried to define the features in a way that they possibly exhibit high values in case of suspicious behavior and lower values otherwise. In this way, when used later on as input for the model, each input feature can directly be interpreted as a key figure for the riskiness of a company with respect to fraud. Hence, this strategy again facilitates our aim to finally achieve an interpretable model that gives the expert the possibility to easily analyze the reasons for specific risk estimations. Because of those considerations during the feature preselection process, we focused on checking the features' suitability as such risk indicators. One major assistance to do so was a histogram visualization as shown in Figure 4 for the example of the average value of the staff's fluctuation over the whole observation period. Besides the usual histogram display with the bin counts, Figure 4 contains additional, distributional information about the target value for each bin. More precisely, the background of each bin is colored with respect of its proportion of unlabeled and suspected companies. Considering all companies belonging to the bin under investigation, the proportion of unlabeled companies is colored in blue, the proportion of suspected companies is colored in red. For example, in Figure 4, 270 companies fall in the second bin which includes the very low indicator values of the interval $[0.005, 0.015]$, only one company thereof (i.e. only 0.37%) is already suspicious. Hence, the background bar is almost completely colored in blue. On the other hand, 5 out of 11 companies of the interval $[0.22, 0.23]$ with high indicator values are on the list of suspicious enterprises, the proportion of 45.45% is displayed with the red part in the background of the bin.

If such a plot, for increasing indicator values, shows a positive trend in the proportion of *suspected companies* (like in Figure 4), the feature can be considered as useful for a subsequent model generation.

Figure 5 shows two additional plots that can help to judge a feature's potential as an indicator. On the left side, there is a box plot that describes the feature's distribution for each target class. For the staff's fluctuation, this box plot clearly exhibits that the mean value of the class *suspected companies* is far higher than for the *unlabeled companies*, even if the two distributions overlap. Hence there is significant potential of the feature for the use as indicator.

The plot on the right side displays how a simple scoring model that uses only the observed single feature would perform. More precisely, first, the feature is sorted with the highest value at the top and the resulting ordered list of the according companies is taken as recommendation list. As mentioned before we aim at obtaining a scoring list that contains "as many suspicious companies as possible" at the top. Therefore, in our case study, a major performance measure for our models was the proportion of already labeled *suspected companies* for any given length of scoring list *ntop* (i.e. taking the top *ntop* entries of the scoring list). The right hand plot of Figure 5 shows the performance curve (blue line) over an increasing list length (starting with $ntop = 1$) for the recommendation list based on $Mean(Fluct)$ evaluated for companies of the construction sector. The

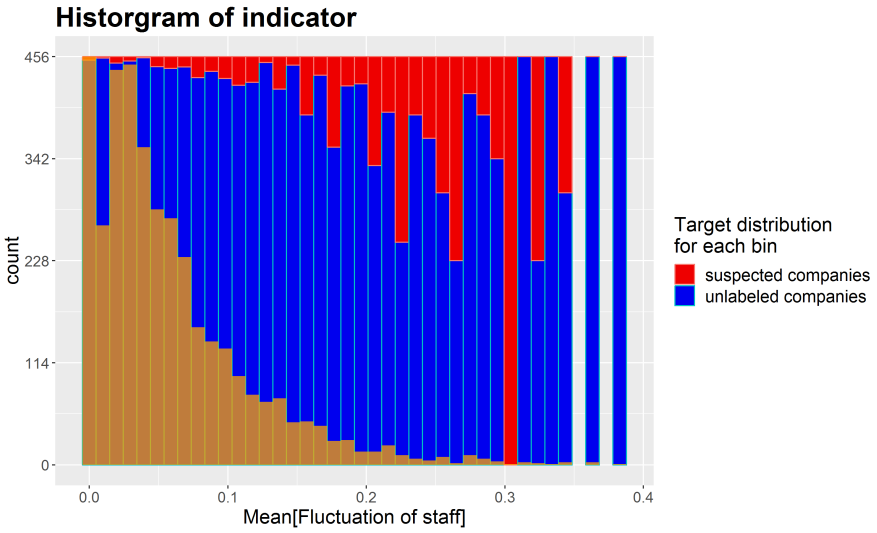

Fig. 4: Histogram of indicator *Mean[Fluctuation of staff]*: colored background shows the target distribution within each bin

red line represents the base line, i.e. the overall proportion of *suspected companies* in the considered data set (for the given example this is 2.76%). During our evaluations, a typical choice for *ntop* has been 40, as this is a reasonable amount of companies that in-house experts consider manageable to handle (for closer investigations per month). For the given example, 11 of the top 40 companies are already *suspected*, that makes 27.5%. Hence, compared to selecting companies by chance (2.76%), choosing the companies according to the indicator $Mean(Fluct)$ (27.5%) results in an information gain by a factor of 10.

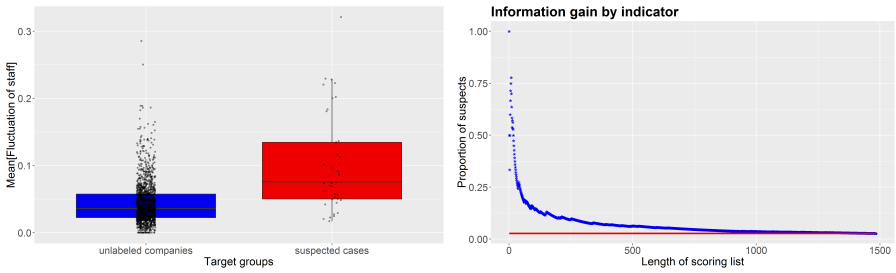

Fig. 5: Plots illustrating the suitability of feature $Mean(Fluct)$ as key figure: a box plot describing the feature's distribution for each target class (left side) and a performance plot of the recommendation list solely based on the feature $Mean(Fluct)$ (right side)

4.3 Model Generation and Selection

As mentioned before, when working on a recommendation model, our focus was on providing the end user, i.e. the in-house experts of the federal health insurances in Austria, with an easily comprehensible model. The set of feature candidates was generated with the intention that each feature is positively correlated with the riskiness of fraud. However, instead of directly taking the features as they are, we rather want to make use of their inherent ranking. More precisely, for each feature the following preprocessing step is conducted: First the feature is sorted with the highest value at the top and from the resulting ordered list the first 500 entries are selected. To these 500 companies there is assigned a linearly decreasing score value from 1 to 0, i.e. the company with the highest feature value is assigned the score 1 (i.e. 500/500), the second highest takes 499/500, and so on; up to the 500th company with 1/500. Finally, all the other companies with even lower feature values obtain score 0. The in this way calculated scoring variables are eventually the figures that are used for the subsequent model generation.

The motivation for this kind of feature preprocessing was twofold: First, aiming for a recommendation list, we are rather interested in predicting the correct inherent ranking of the companies' risk for fraud than estimating the exact risk for every single company. Therefore we expect that using the ranking information instead of the features' absolute values is the more appropriate way to go. The second benefit we see is that with this feature treatment every single indicator automatically has the same distribution. Hence, all the feature values are directly comparable what will significantly ease the generation of a well-balanced model as well as its interpretability.

At least in the first stage, we decided to stick to linear models. Together with the properties of the deduced features, linear models allow that its coefficients can directly be interpreted as the corresponding feature's contribution to the risk estimation. However, first modeling attempts demonstrated that regression models trained on the available data including the whole generated feature set tend to focus on only a few, dominant subgroups of the known suspicious patterns covered by the available features. However, an important requirement from the experts' side has been the broadness of the resulting model, i.e. the recommendation list shall contain cases that preferably cover multiple (as many as possible) aspects of suspiciousness at the same time.

Therefore, the following two-step approach was realized: First of all, all available indicator variables are manually divided up into six different context groups. For example, all indicators dealing with the payment history of an employer were put together to the indicator group *financial abnormalities*, all the features referring to the development of the staff (like $Mean[Fluct]$ from the previous subsections) form the indicator group *abnormalities of staff's fluctuation*. Based on this additional context information we are aiming for a model that contains reasonable contributions from all indicator groups.

To ensure this, in a first step, models are generated for each indicator group individually. That means that, for each specific indicator group, an extensive grid

search is conducted trying out all possible combinations of features. Considering a maximum length of 10 features for a single indicator group, we end up with up to 1023 different models per group ($2^{10} - 1 = 1023$ is the total number of subsets for a set of 10 features). Next, all these models are evaluated by calculating the proportion of *suspected companies* in the according recommendation lists. Additional to the set of all companies of the economic sector under investigation this performance evaluation is always repeated twice on subsets of the data restricted by the company size. The reason for this is that the in-house experts sometimes want to focus their search on enterprises with comparable company size. Thus, the chosen model needs to perform well also in these settings. Finally, based on performance figures as well as other fuzzy criteria (like interpretability, number of used features, ...), up to 5 possible feature combinations are chosen for each indicator groups.

In a second step, a further grid search is conducted to achieve final model candidates, this time combining the in step one identified feature combinations from the different indicator groups. In this way we can guarantee that all indicator groups are represented in our final model and reducing the model with respect to a specific aspect (i.e. measuring the contribution of a certain indicator group) will still result in an optimal model for the according restricted setting. The evaluation and final selection of the model were done equivalently to step one.

5 Results and Experiences

The approaches presented in the previous sections were developed in collaboration with the Upper Austrian health insurance where, in a first stage, we worked on models for the economic sectors of construction and gastronomy based on data of the federal region of Upper Austria. The differentiation of the economic context was motivated by the well-known fact that between the sectors relevant characteristics may differ a lot. Highly suspicious patterns in one sector might express totally normal behavior for another sector. Even when focusing on single economic sectors, we eventually had to deal still with a high heterogeneity of the data, like different economic subsectors or differing scales of company size, to name two major sources.

We had to make the same experiences when we tried to directly transfer the models for Upper Austria to the other 8 federal regions of Austria. Even if one can expect that all in all most of the suspicious patterns somehow exist in all the regions, the differences (be it occasional regionally isolated patterns and/or different local expressions of general patterns) apparently were too big to achieve satisfactory results when simply applying the models developed based on the data of Upper Austria. Thus, we ended up in carrying out the same procedure as described in Section 4 for each region and both the economic sectors. That means that we have developed and included to the RAD-Tool altogether 18 recommendation models that are currently in use at the regional health insurances.

With respect to the models' performance, we can make the qualitative statement that "from a statistical point of view the models perform well", in the sense that the proportion of suspicious companies in the resulting recommendation lists is significantly higher than the chance levels. Furthermore, the feedback of the in-house experts that are using the outcomes of the models is mostly promising. However, evaluating our results in a quantitative manner remains a pending task. The reasons for it are manifold, most based on the fact that the output of the models is not directly processed, but only supports the further decision-making process of the human expert:

- The benefit of the model is not only defined by the prediction accuracy, but many fuzzy, difficult to measure criteria (like interpretability of found cases or possibilities of legal prosecution) play a role.
- The financial benefit, eventually the crucial, most interesting figure, is extremely difficult to measure. Effective financial results can often be realized only after a long legal process (months or even years). And even then a quantitative estimation is not always doable.
- Moreover, it is difficult to estimate the exact contribution of the recommendation lists as there are many different factors that lead to a specific financial result.
- Finally, the available amount and quality of target data (i.e. the list of *suspected companies*), at the moment, is not good enough to perform a thorough evaluation.

6 Conclusions and Future Work

In the context of social insurance, it is very important that the available resources might be devoted, and to a greater extent, to those who really need them and pursue any situation in which public funds are used for an unintended purpose. With the aim of providing methods and tools to do that, we have presented here our data-driven approaches for fraud detection in the context of employers that do not follow the rules for hiring workers focusing the competition on the lowest price.

In the future, we want to work towards a fully automated, comprehensive model that is applicable to all of Austria and different economic sectors simultaneously. Benefits of such an all-in-one model would be

- comparability over regions and economic sectors,
- lower costs for quality assurance, and
- exploitation of useful information over multiple related tasks.

We plan to go in this direction by the use of multi-task learning, possibly in combination with learning to rank as well as methods that fit well for partially unlabeled data (to address further particular characteristics of this use case's data).

Still another promising direction will be the use of unsupervised learning methods. For example, clustering of the data and subsequent proper description

of the interesting clusters could lead to interpretable models. Additional advantages could be better manageability of the data's heterogeneity as well as the potentiality of discovering even unknown suspicious patterns.

Acknowledgments

The research reported in this paper has been supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center SCCH.

References

1. M. Diaz-Granados, J. Diaz Montes, M. Parashar: Investigating insurance fraud using social media. BigData 2015: 1344-1349.
2. P. Dua, S. Bais: Supervised Learning Methods for Fraud Detection in Healthcare Insurance. Machine Learning in Healthcare Informatics 2014: 261-285.
3. M. M. Hofmarcher, W. Quentin: Austria: Health system review. Health systems in transition 15.7: 1-292 (2013).
4. R. M. Konijn, W. Kowalczyk: Finding Fraud in Health Insurance Data with Two-Layer Outlier Detection Approach. DaWaK 2011: 394-405.
5. I. Kose, M. Gokturk, K. Kilic: An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Appl. Soft Comput. 36: 283-299 (2015).
6. F. Lu, J. E. Boritz: Detecting Fraud in Health Insurance Data: Learning to Model Incomplete Benford's Law Distributions. ECML 2005: 633-640.
7. B. Rao: The role of medical data analytics in reducing health fraud and improving clinical and financial outcomes. CBMS 2013: 3.
8. V. Rawte and G. Anuradha: Fraud detection in health insurance using data mining techniques, 2015 International Conference on Communication, Information and Computing Technology (ICCICT), Mumbai, 2015, pp. 1-5.
9. Y. Shi, Y. Tian, G. Kou, Y. Peng, J, Li: Health Insurance Fraud Detection. In: Optimization Based Data Mining: Theory and Applications. Advanced Information and Knowledge Processing. Springer, London (2011).
10. C. Sun, Q. Li, H. Li, Y. Shi, S. Zhang, W. Guo: Patient Cluster Divergence Based Healthcare Insurance Fraudster Detection. IEEE Access 7: 14162-14170 (2019).
11. D. Thornton, G. van Capelleveen, M. Poel, J. van Hillegersberg, R. M. Mller: Outlier-based Health Insurance Fraud Detection for U.S. Medicaid Data. ICEIS (2) 2014: 684-694.
12. Y. Tsai, C. Ko, K. Lin: Using CommonKADS Method to Build Prototype System in Medical Insurance Fraud Detection. JNW 9(7): 1798-1802 (2014).
13. V. Van Vlasselaer, T. Eliassi-Rad, L. Akoglu, M. Snoeck, B. Baesens: GOTCHA! Network-Based Fraud Detection for Social Security Fraud. Management Science 63(9): 3090-3110 (2017).
14. A. Widder, R. von Ammon, G. Hagemann, D. Schoenfeld: An Approach for Automatic Fraud Detection in the Insurance Domain. AAAI Spring Symposium: Intelligent Event Processing 2009: 98-100.
15. W.-S. Yang, S.-Y. Hwang: A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst. Appl. 31(1): 56-68 (2006).