Medi
Maybe
INSURANCE

Fraud Detection Assessment Report

for *MediMaybe Insurance Company*

Prepared by: Jan Bohnenstengel, TryingOutConsulting
(employee-ID: h12426307)

Scope: Reassessing Claim Processing

Submitted to: Florian W. Szücs, CEO

June 22, 2025

Executive Summary

This report assesses MediMaybe's fraud detection performance and proposes actionable improvements using data-driven methods.

Key Findings:

- MediMaybe currently underperforms industry benchmarks in fraud detection. Only 6.26% of fraudulent claims are correctly flagged (recall), compared to an industry-estimated fraud rate of 5–10%.
- 93.74% of fraudulent claims go undetected, costing the company approximately **63.3 million EUR** in unnecessary settlements over the observed period.
- The current rule-based classification system relies on several criteria, but only three of the eleven rules produce meaningful separation between fraud and non-fraud. Many rules lack statistical significance.

Regional and Temporal Patterns:

- Smoothed fraud rates vary significantly across ZIP-code-defined regions, with some regions displaying nearly double the fraud rate of others.
- Temporal trends show peaks in both claim volume and fraud rate during 2008–2010, indicating possible systemic or seasonal effects.

Proposed Improvement: Predictive Modeling

- A gradient boosting model was trained using historic claim data (excluding settlement outcomes) and calibrated to match current specificity.
- At the same false-positive rate, this model improves recall to 9.24% and reduces fraud settlement cost by over **7.3 million EUR** — a 12% reduction without increasing customer complaints or reputational risk.
- The model identifies financial variables, internal recommendations, and timing metrics as key predictive features.
- SHAP analysis confirms model transparency and allows operational insights into fraud patterns, including high-impact variable combinations.

Recommendation: Introduce predictive modeling to support or replace the current rule-based system. A simple, interpretable model already offers measurable gains, and further improvements are likely with additional data or modeling refinement. Thresholds can be tuned to balance direct cost savings with reputational risk, ensuring continued customer trust while reducing fraud losses.

Current State of Company and Objectives

Metric	MediMaybe Insurance	Industry Average
Revenue per policy	70	100
Contribution from core operations	20 %	35 %
Contribution from investments	80 %	65 %
Claims settlement ratio	92 %	87 %
Claims repudiation ratio	2.5 %	4 %
Average settlement period (days)	72	30–45
Net promoter score	73	86

Table 1: Company Performance vs. Industry Average

MediMaybe is currently underperforming the industry average in every relevant metric (See Table 1). While the revenue per policy is already quite low, the settlement to claim ratio is relatively high, implying low claims repudiation as shown with 2.5 % in the table. This is significantly lower than the industry average of 4 %, estimates of actual fraud rates from the US health insurance industry even range from 5-10 % (Jacobs, 2007). Thus, there is likely a lot of money spent to settle claims, which in reality are fraud. This is particularly a problem in this company, as it relies heavily on its core business, health insurance, compared to other investments. Therefore improving the detection of fraudulent cases is an imminent and necessary step to take. However, just generously flagging more claims as fraud, while detecting more actual fraudulent claims, of course increases false flaggings, accusing innocent customers of fraud. This would lead to an increasing loss of reputation, which would be devastating for a company relying on core business and a high volume of cases/clients, as the revenue per policy is quite low as described. A connected problem is a long processing time of claims, making people wait for a decision, certainly not benefiting public opinion either. As the net promoter score is already below industry average, this is important to be kept in mind.

Therefore, I was given the task to reassess 80,000 historical claims from between December 2004 and December 2011, processed by MediMaybe Insurance, to draw conclusions on the current performance in fraud detection, patterns in the data, and how to make improvements to increase profitability of the firm. Special attention will be given to the tradeoff between identifying more fraudulent cases, reducing unjust settlement cost, and not accusing innocent customers of fraud, to prevent a loss in reputation and possible compensation or legal cost. Fortunately, every claim has been assessed by the best expert consultant on insurance fraud and labeled fraud or non-fraud, which supports this analysis significantly.

Current Classification Performance

Given the expert classifications, it is possible to asses the current company performance in fraud detection very clearly. To assess the performance of classifications, 2 measures will be used:

1. **Recall**: amount of rightfully detected frauds as share of amount of actual frauds (share of true positives, TP). A higher recall reduces paying out settlements for claims that are actually fraudulent (false negatives, FN).
2. **Specificity**: amount of rightfully detected non-frauds as share of amount of actual non-frauds (share of true negatives, TN). A higher specificity reduces the number of false accusations of fraud, which might lead to loss of reputation or compensations/legal cost (false positives, FP).

First, it has to be noted that that the share of fraudulent claims at MediMaybe Insurance is excessively high, namely at 21.7 %, which is more than double the upper bound of the estimated industry average named above. Now to the performance of classifying so far (Table 2): The current system classifies cases as *genuine*, to *discuss* or to *investigate*. Declaring the labels *discuss* and *investigate* as declaring a claim fraud, only 6.26 % of frauds have been rightfully detected (= recall). This implies, that 93.74 % of actual frauds have been missed and settled, amounting to 63,270,740 EUR paid out in settlements of fraudulent cases. Considering the total settlements amounting to 313,724,610 EUR, this is a highly significant share. The current classification even predicts a higher share of frauds among actual honest claims than among actual fraud claims, so there is a lot to improve.

	Predicted Fraud	Predicted Non-Fraud	Share classified fraud
Actual Fraud	1,086 (TP)	16,274 (FN)	6.26 %
Actual Non-Fraud	5,348 (FP)	57,292 (TN)	8.54 %
Share actual fraud	16.88 %	22.12 %	
Recall (Fraud): 6.26%	Specificity (Non-Fraud): 93.74%		

Table 2: Model Confusion Matrix and Key Metrics

The share of false positives, so false accusations of fraud, is luckily quite low. When trying to improve the classification this has to be taken into account. Increasing recall often decreases specificity, which is only beneficial under certain conditions and depends on the relative costs of false positives vs. false negatives. This trade-off will be discussed later, when possible improvements are presented.

Now to the procedure of current classification. Like many other companies in the industry, MediMaybe uses a rule-based scoring system to classify claims as possible frauds. The rules, which I define as binary, are the following:

1. Claim from non-networked hospital? - Yes/No

2. Multiple Claims from a single policy? - Yes/No
3. Multiple claims from same hospital on the same day? - Y/N (threshold >= 3)
4. Claim close to policy expiry? - Yes/No (threshold 14 days before)
5. Claim before/close after policy start? - Yes/No (threshold 30 days after)
6. No pre- or post-claims? - Yes (= no pre-/post-claims)/No
7. Misrepresented information? - Yes/No
8. Claim on weekend? - Yes/No
9. Costly expenses? - Yes/No (to 5 % of any expense group)
10. Claim earlier than 1 day before discharge? - Yes/No (actually 11)
11. Claim later than 2 days after submission? Yes/No (actually 12)

For every question answered with yes, a point is assigned. The overall score is then a weighted sum of these points. The maximum score is 40. If the score is lower than 15, the claim is automatically labeled *genuine* and not further investigated. For a score of 16-20 the claim is labeled *discuss*, and more additional data is requested from the information collection team and the healthcare provider. For a score above 20, the claim is labeled *investigate*, and forwarded to fraud investigation team. In this analysis for simplicity the labels *discuss* and *investigate* are usually seen as predicted frauds *investigate* as described above.

To assess this system's performance more detailed, Table 3 summarizes an analysis of the power of each of these rules. For each the sample is divided by the value of the rule variable, then computing fraud rates for each subgroup. The differences between these means is displayed and then tested for statistical significance via t-tests.

Variable	No	Yes	Difference
Rule 1 (non network)	0.2156	0.2184	0.0029
Rule 2 (multiple claims)	0.2171	0.2162	0.0009
Rule 3 (group claim)	0.2174	0.2166	0.0008
Rule 4 (near expiry)	0.2190	0.2092	0.0098***
Rule 5 (near start)	0.2172	0.2151	0.0021
Rule6 (no prepost)	0.2457	0.2160	0.0297***
Rule 7 (misrepresentation)	0.2126	0.2187	0.0060*
Rule 8 (weekend)	0.2174	0.2159	0.0015
Rule 9 (costly)	0.2174	0.2155	0.0020
Rule 10 (day before discharge)	0.2162	0.2183	0.0021
Rule 11 (claim after 48h)	0.2163	0.2185	0.0022

Table 3: Fraud Rates by Rule-Based Flags. Stars show significance based on t-tests. * = 10 %; ** = 5 %; *** = 1 %

The only rules that lead to significant differences are the rules 4, 6 and 7. The highest difference is for rule 6, which is triggered when there are no claims for pre- or post-hospital expenses. The fraud rate of such claims is at 24.57 % and thus 2.97 percentage points larger than that of claims with pre- and/or post-hospital claims. This difference is statistically significant at the 1 % level. Another rule which seems to work is whether the claim was made closer than 15 days to the policy's expiry. Here the difference in fraud rates is only 0.98 percentage points, but still significant at the 1 % level. The last rule is rule 7, indicating misrepresented data. It is triggered for example by negative values for any numeric variables, zero-values for claim or insured sum, or senseless pairings like hospital expenses without a hospital stay or a claim without any medical charges. This leads to fraud rates of 21.26 % for claims without misrepresented data and 21.87 % for those where the rule applies, resulting in a difference of 0.6 percentage points, which is however only significant at the 10 % level. All the other rules seem to have no significant impact when splitting the sample, explaining the poor performance of this classification method.

To dig deeper, the most common pairs of rule violations for cases which were fraudulent were checked. They are the following:

1. Rule 6 (no prepost) & Rule 7 (misinformation): 12162 cases
2. Rule 3 (group claim) & Rule 6 (no prepost): 8896
3. Rule 1 (non network) & Rule 6 (no prepost): 8092
4. Rule 3 (group claim) & Rule 7 (misinformation): 7192
5. Rule 11 (day before discharge) & Rule 7 (misinformation): 6316

We see again, that rule 6 and 7 appear often. A logical extension of the rule based approach would be to check for combinations (comparable to interactions) as well.

Overall, it is clear, that this system's performance is not great at detecting fraudulent cases at all. On the way to find possible improvements, in the next section, the data is scanned for patterns or clusters in terms of claim or fraud behaviour.

Data Analysis

Developments Over Time

Even though there are cases starting 2004 up to December 2011, the period where significant numbers of claims are available are December 2008 until March 2011. Figure 1 displays the monthly number of claims as well as the monthly fraud rate over that time frame, as especially the fraud rate is not worth interpreting in the other months with usually a low single digit number of claims. The number of claims peaks twice, first in the first half of 2008, before declining over the year 2009 and than spiking hugely in the first half of 2010 again, with a maximum of around 7100 claims in April 2010. Afterwards the number of cases decreases steeply again. Fraud rate seems to move in a similar pattern, just smoothened strongly. In the first period with many claims it peaks around 22.5 %, before declining to

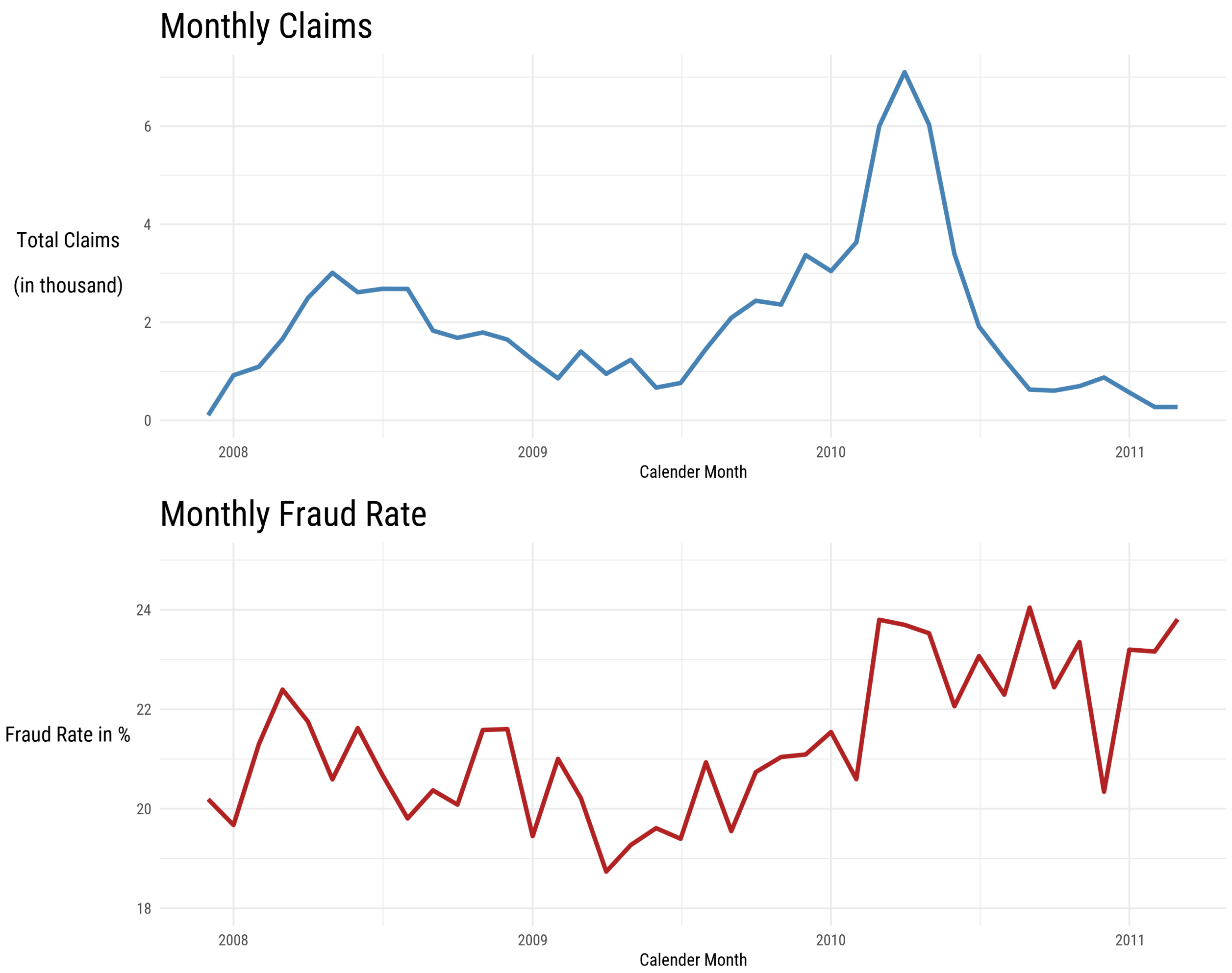

Figure 1: Claim amounts and fraud rate over time.

approximately 19 % in the spring of 2009, where the number of claims plummits as well. Around spring of 2010, where the number of claims is high as well, the fraud rate goes up to 24 %, before the number of cases drops so far that it's not interpretable anymore. This suggests a possible correlation between the monthly number of claims and the fraud rate.

Age Groups

Age Group	Total Amount of Claims	Total Number of Frauds	Fraud Rate
<25	21911	4718	0.215
25–39	23663	5131	0.217
40–54	17395	3734	0.215
55–69	14097	3117	0.221
70+	2932	659	0.225

Table 4: Comparison of key metrics across age groups

Table 4 gives an overview over key metrics across age groups. Most claims are submitted by people between 25 and 39 years old. This seems slightly counterintuitive as the median age in most developed nations is around 39 years or higher and usually required medical care increases at older age, suggesting more claims from that group. This could just be from sampling or the specialization of this company on certain medical services.

The fraud rate does not vary much, however, the seems to be a possible discontinuity at age 55. A t-test however finds the difference of the average fraud rates for the general age groups under 55 and 55 or older to not be statistically significant at any usual level. Grouping by age seems to result in no patterns at all.

Regional Level

The analysis at regional level required some data preparation. Regions were defined by ZIP codes, which are a systematic combination of letters and numbers in the following way (A representing a letter, 1 a number): AA1(1)A(A)1(1). All ZIP codes starting with the same 2-letter prefix were assigned to the same region. As here the same problem of few claims per region appears again, the usual fraud rates are not of great use. To create an interpretable metric, fraud rates are smoothed, similar to Bayesian methods. The first one is computing the fraud rate at ZIP code level, adding 50 fictional claims with their value for fraud being the overall average fraud rate as a "prior". This is a rather strong prior, given the number of 50 fictional claims, smoothing the rate strongly towards the overall average fraud rate, being particularly robust to small claim numbers. At region level the ZIP level smoothed rates are then just averaged. The second metric is computed at region level directly with a weaker prior, using only 5 fictional claims. This way, this second smoothed fraud rate measure picks up more anomalies, but is more prone to distortion by small samples. The whole data set was then collapsed to region level, to allow comparison of useful metrics at this level.

Figure 2 compares these regions across the smoothed rates. It shows the 10 regions with highest and lowest smoothed rate with strong prior each. Depending on the prior strength and level of smoothing the rates and even the ordering of regions differs, but the usage of rates depends on the objective. Generally, it shows strong differences among regions exist. Even the rather strongly smoothed rate, which was smoothed at ZIP level, ranges from around 18 % to 33 %. The rate with weaker prior even ranges from 4 % to 46 %. There is no clear trend for the fraud rates with respect to the amount of claims within a region. Excluding regions with enormous amounts of claims (> 1000), which are only 4, but have smaller fraud rates, there is a small positive relationship, but it is small and it is possible to have been cased by smoothing the rates. Therefore this can not really be considered. The differences between regions still persist. With additional information on the regions, e.g. their geographical location, it could be possible to find clusters of regions.

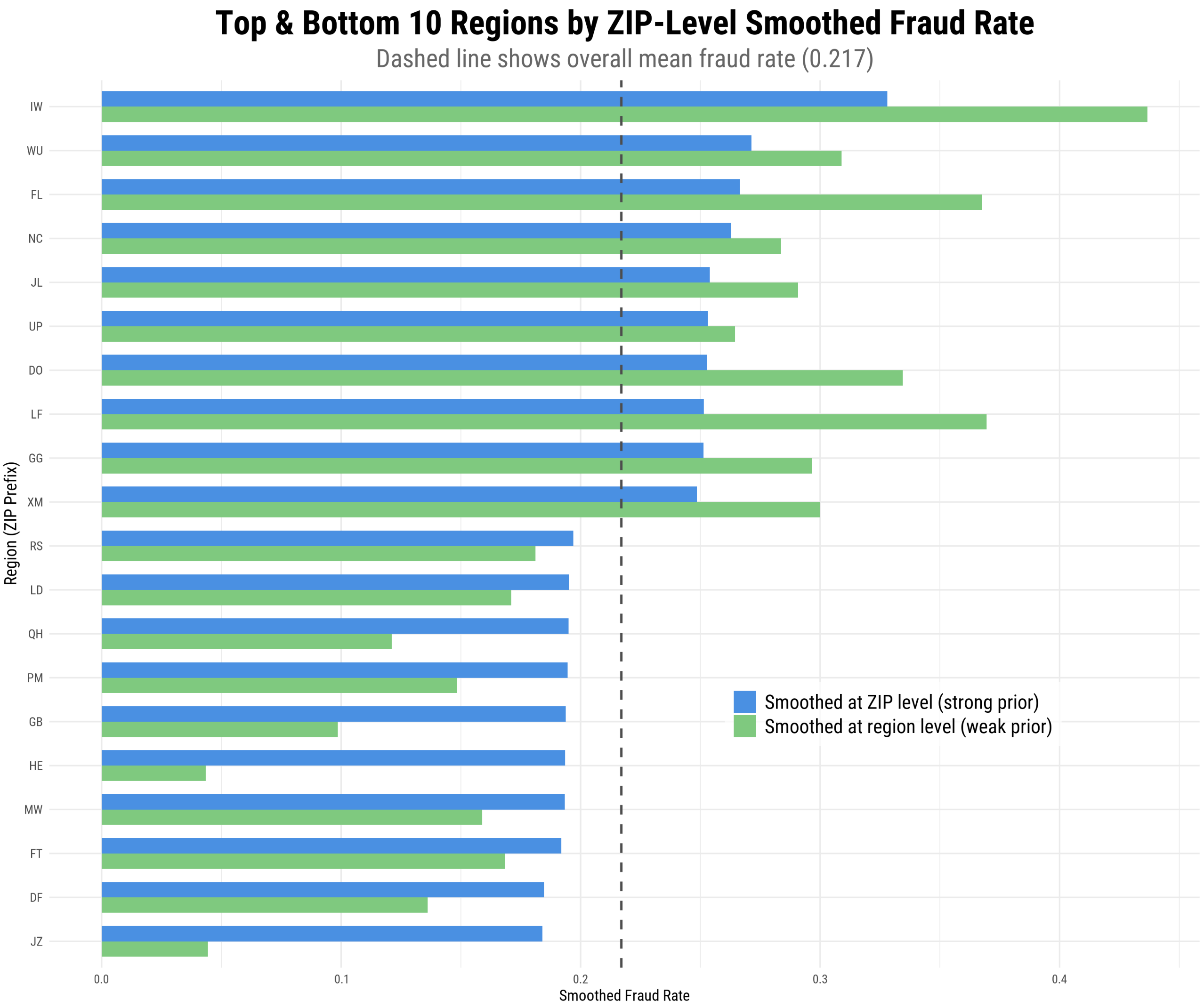

Figure 2: Smoothed fraud rates across regions.

How to improve

There are several ways to improve the performance. One way that is quick to implement and use patterns in the data, even at higher order correlations, non-linear relationships and interactions, are predictive models. Especially with the trend of increasing availability of data they become more and more useful. Machine learning models can use vast amounts of data and can outperform traditional classifying methods like the rule based method used in this company. The use of such methods is also highly flexible. Predictive models can just be added into the current procedure of processing claims, a model could only select other variables to increase the predictability of the rule-based scoring method or replace that completely. In the case of better performance the processing duration can be reduced as well. They are also flexible in calibration, there is a vast selection of methods, and for each one parameters can be adjusted. In addition other rising, but yet not widely used methods like text or image analysis could complement that (InsuranceFraud.org, 2025).

To provide a concrete example, I created and trained a predictive machine learning model on the data I was analyzing. I used a gradient boosting model using five-fold cross validation, as it balances strong predictive power with still a degree of interpretability, by reporting variable importance for example. To simulate the situation of the model facing new claims, I removed all data about settlement timings, amounts or further information only available after a decision about fraud by the company would be made. Also I kept the model rather simple, adding time variables like time between admission to a hospital and the submission of the claim, but no extensive variables on policy level. By including more variables, e.g. about the claim history policies, the model could surely always be improved further.

As the costs of false positives is unknown to me I first tried to calibrate the model so it resembles the conventional rule-based approach in specificity. This was done when I set the threshold for which predictions are classified as fraud to > 0.5541. As can be seen in Table 5, by that the specificity becomes 91.46 %, so exactly that of the rule-based approach, but the **recall increases to 9.24 % compared to 6.26 %** of the conventional approach. The cost for settlements paid to claims which are in reality fraudulent (cost of false negatives) decreases to 55,979,505 EUR. Compared to the 63,270,740 EUR from the method used so far, this is a **decrease of 7,291,235 EUR or 12 % in fraud payouts, without increasing false-positive flags**. This means a big decrease in certain direct cost without increasing indirect cost such as compensation, legal cost or very importantly fewer clients due to loss of reputation, by accusing more innocent people of being fraudulent. For the average truthful customer the experience with the company would stay the same. And all that just with a rather simple model.

Model	Recall	Specificity	FN	FP	Cost of FN (EUR)
Benchmark method	6.26%	91.46%	16,274	5,348	63,270,740
Gradient Boosting (p > 0.5541)	9.24%	91.46%	15,756	5,348	55,979,505
Gradient Boosting (p > 0.560)	7.51%	93.10%	16,057	4,322	57,205,730
Gradient Boosting (p > 0.56628)	6.26%	94.46%	16,274	3,468	58,131,090

Table 5: Model Performance Comparison

Depending on the cost of false positives, decreasing their rate can be a major concern as well. This is of course possible, too. At a threshold of > 0.56628 the gradient boosting model matches the rule-based scoring's recall (i.e. the number of FN), but then specificity is increased to 94.46 %. This means, **keeping the fraud detection rate constant, 1180 claims or 35 % in relative terms less would be falsely flagged as possible frauds**. Interestingly, even though the number of FN is the same, the total amount paid in settlements for FN is still lower than that of the conventional predictions. The model apparently not only detects more, but also the more expensive frauds. Any threshold between those leads to a scenario with the measures between those scenarios above. Therefore, depending on the relative size of cost of FP and cost of FN the threshold can be adjusted accordingly. Overall performance can be increased by including more variables or calibrating the model more thoroughly. Also other model classes are thinkable.

In addition, the model can tell, which variables is used to what extend to make predictions. A very helpful to here is SHAP-analysis. SHAP (SHapley Additive exPlanations) is a model-agnostic way to asses the importance of variables for predictions of machine learning models. SHAP values measure how much each feature contributes to shift the prediction away from being non-informative. I.e. a high negative value implies that the feature pushed the probability of being fraud significantly towards 0, while a high positive value pushes towards 1. However, this can not just be interpreted as what this variable does independently of anything else. SHAP measures the marginal contribution of a feature averaged over all

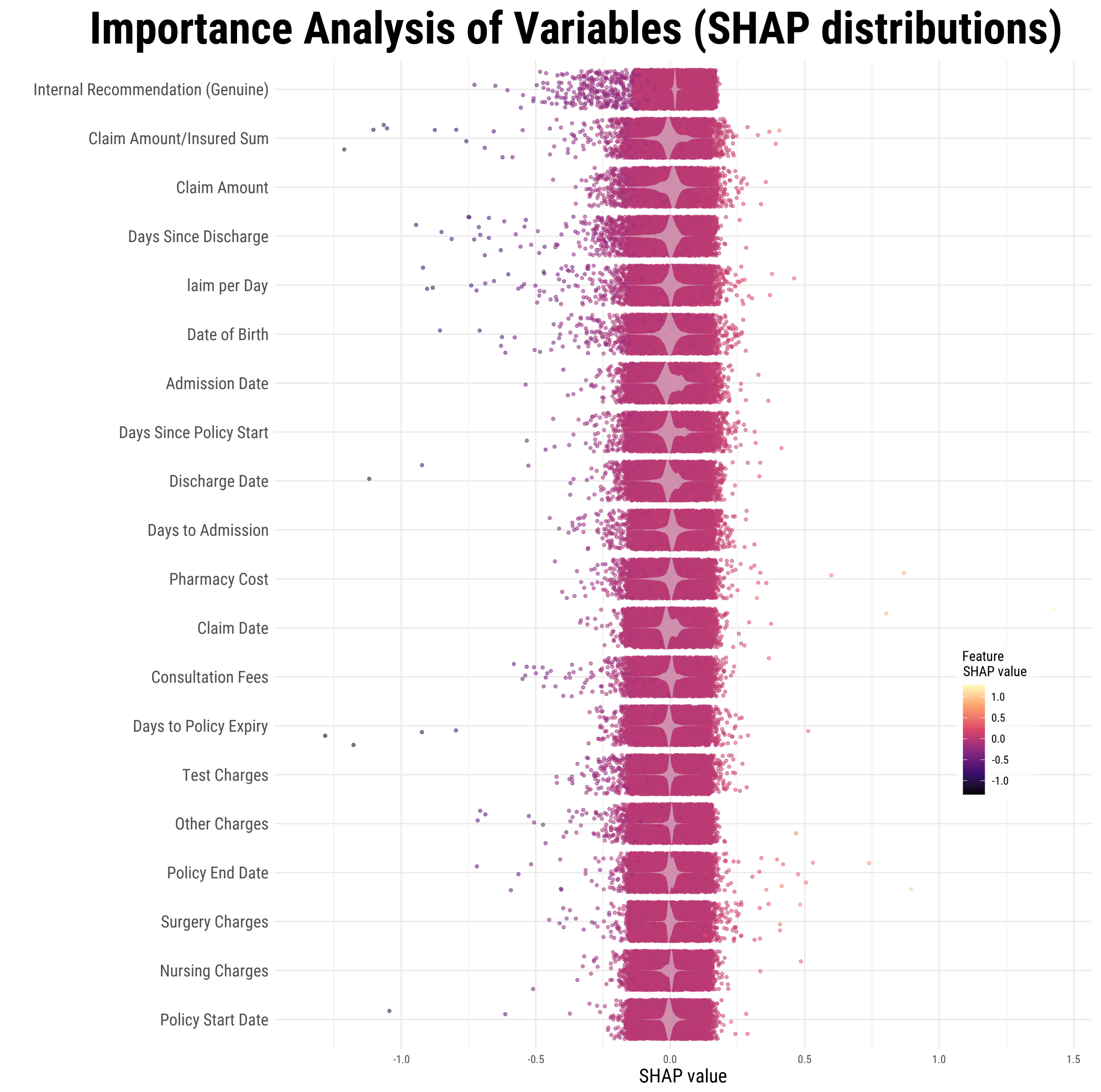

Figure 3: Variables Ranked by Importance for the Gradient Boosting Model (SHAP Analysis)

possible contexts. So even with high SHAP values a variable itself might be almost useless, but in combination with others it may become very strong and therefore receive high SHAP values.

Having this in mind, Figure 3 displays the the SHAP values of the 20 variables used in the gradient boosting model, which on average give the highest contributions to the model's predictions. Surprisingly, even though its direct predictable power was so bad as seen in the section about the conventional rule-based classification's performance, in the machine learning model the recommendation variables contributes the most to the model. Also financial variables like the claim amount, the ratio of claims to the insured sum or individual charge variables influence the predictions significantly. Time-based features like days between discharge from the hospital and submission of the claim are are important as well. Interestingly date of birth being there hints towards age having an impact, even though there was no obvious effect found in the data analysis section. As mentioned this is likely due to interactions.

Overall it can be said, that including predictive models in the company's toolbox would be an easy way to improve. It is easily applicable, flexible in use, increases performance drastically and can also shorten processing time for claims. The danger of increasing false positives can be remedied by adjusting parameters in the model as shown.

External Sources

InsuranceFraud.org. (2025, April 27). Fraud Stats - InsuranceFraud.org. InsuranceFraud.org -. https://www.friss.com/insurance-fraud-report-2022

Jacobs, A. (2007, July 1). Challenge facing payers is to reduce prevalence of healthcare fraud. Managed Healthcare Executive. https://www.managedhealthcareexecutive.com/view/challenge-facing-payers-reduce-prevalence-healthcare-fraud