# A Prescriptive Machine Learning Approach for Assessing Goodwill in the Automotive Domain

Stefan Haas✉[1] and Eyke Hüllermeier✉[2]

[1] BMW Group, Munich, Germany, stefan.sh.haas@bmwgroup.com
[2] Institute of Informatics, University of Munich (LMU), Germany,
eyke@lmu.de

**Abstract.** Car manufacturers receive thousands of goodwill requests for vehicle defects per year. At BMW, these requests for repair-cost contributions are either assessed automatically by a set of fixed rules or manually by human experts. To decrease manual effort, which is still around 50%, we propose a machine learning approach with the goal to discover so far unknown assessment patterns in human decisions. Since the assessment contribution data is heavily imbalanced, we structure the learning task hierarchically: The first layer's task is to predict the main rank of the request (no contribution, partial contribution, or full contribution). Then, in the case where partial contribution is suggested, the second layer predicts the concrete percentage using a regression model. To optimize our model and tailor it to certain strategies (e.g., customer friendly or more cost oriented), we make use of a custom-defined cost matrix. We also outline how the model can be used in a scenario in which it prescribes appropriate monetary contributions for requested repair-costs. This can initially happen in the form of a decision support system (DSS) and, in the next step, through automated decision making (ADM), where a certain part of goodwill requests is processed automatically by the prescriptive model.

**Keywords:** Prescriptive Machine Learning· Decision Support Systems · Automated Decision Making · Cost-Sensitive Learning· Hierarchical Learning

## 1 Introduction

Rule-based expert systems are used widely in many fields, for example in industry to assess financial credit risks or in medicine to detect diseases such as breast cancer or diabetes [1, 8]. They arguably constitute the simplest form of artificial intelligence (AI), storing rules carefully assembled by domain-knowledge in the form of if-then-else statements. They do not require any data and are *naturally interpretable* [2]. This makes them a natural fit for automating decision processes that need to be auditable, 100% accurate, and which comprise a certain risk, either financially or for life and limb.

One such financial rule-based expert system is the central Goodwill system of BMW. In cases of vehicle defects, dealers carry out goodwill repair on behalf of customers and in turn get compensated by the original equipment manufacturer (OEM) for their spare parts and labor efforts. Whether or not customers are eligible for goodwill compensation is decided automatically on the basis of a fixed set of expert rules. This automatic rule based assessment is only done in countries where no legal restrictions against it apply. In case the goodwill request is rejected in the first place, the final decision is transferred to a so-called *assessor*, a human after-sales goodwill expert, who manually looks at the individual case and determines the monetary contribution of the OEM, if any. Although a decision matrix to support this manual process is in place in many sales markets, it is still often a commercial gut decision and not standardized across markets.

The need for human intervention is due to several problems of a rule-based approach, notably the difficulty to maintain a coherent set of deterministic rules capturing all eventualities of a complex commercial use case. Therefore, the data-driven design of decision models by means of machine learning (ML) appears to be an appealing alternative to increase the degree of automation. Over the years, a good amount of historic human decision data has been collected, which can be leveraged in this regard. The goal hereby is to deduce so far unknown assessment patterns from observed human decisions that might be too complex to be put into rules in the first place. Supervised machine learning models can be trained on the observed decision data and later used in the manual decision process to *prescribe* certain monetary contributions. This can either happen in the form of a *decision support system* (DSS) or, if trust in the models is high enough, through *automated decision making* (ADM), which helps decrease manual human assessment effort and save costs in the long run.

The goodwill use case qualifies as what has recently been coined *prescriptive* machine learning [7]. In contrast to the common setting of *predictive* machine learning, the goal is not to predict some underlying ground-truth, but rather to learn models that stipulate appropriate decisions or actions to be taken in order to achieve a certain goal (i.e., to answer the question “How to make something happen?” rather than “What will happen?”). In fact, in the case of goodwill, there is nothing like a “right” monetary contribution. Instead, a decision is more or less appropriate, fair for the customer and strategically opportune for the company. Such decisions are supposed to ensure customer satisfaction while remaining economically reasonable from a manufacturer’s perspective. In addition to increasing the degree of automation, prescriptive models may also contribute to the standardization, consistency, and objectivity of the decision process.

The main contribution of this paper is a prescriptive ML approach to goodwill assessment, which is based on real human decision data. In the next section, we describe the goodwill assessment problem in more detail. Next, we outline how prescriptive ML could be incorporated into the existing process. Then, we propose an ML method for goodwill assessments, which is specifically tailored to the use case and properties of the data. Finally, we conclude with related work, identify challenges and outline directions of future work.

## 2 The Vehicle Goodwill Assessment Process

Assessing goodwill requests is an important topic for manufacturers. In case of BMW, dealers yearly submit thousands of goodwill requests for vehicles that must be assessed. The question whether goodwill is granted or not, and which amount, is far from trivial. It is an individual *commercial decision* that must balance customer satisfaction and financial impact. In this regard, it is important to distinguish between *warranty*, which is a legal obligation for manufacturers, and *goodwill*, which is a non-obligatory service manufacturers provide to customers outside the *warranty* time window (usually after 3–5 years). The goal of compensating customers for product failures outside the *warranty* time window is primarily to safeguard customer satisfaction and loyalty with the brand.

At the OEM, handling goodwill on system level is currently a hybrid approach based on automatic and human manual assessment. The UML Use-Case diagram in Fig. 1 depicts the process and its actors.

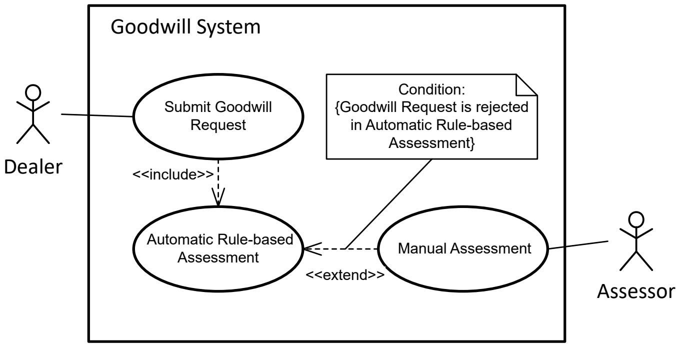
Goodwill System
Dealer
Submit Goodwill Request
Condition: {Goodwill Request is rejected in Automatic Rule-based Assessment}
<<include>>
Automatic Rule-based Assessment
Manual Assessment
<<extend>>
Assessor

Fig. 1: UML Use Case Diagram for the classic goodwill process.

The standard use case is as follows. Customers arrive at a dealership with a vehicle defect and request a repair from the dealer. Next, the dealer checks whether the manufacturer would grant goodwill for this particular defect by submitting a goodwill request on the behalf of the customer. The data the dealer has to enter ranges from certain vehicle information like vehicle mileage and age to estimated labor and parts costs for the repair itself. On system side, the request is first evaluated against a fixed set of rules (automatic rule-based assessment). If it goes through and goodwill is granted, the process is finished and the dealer will be compensated for the repair. If not, the goodwill request is further processed through a *manual assessment*. In this case, a human goodwill after-sales expert checks the request and makes the final decision. The manual assessment step only extends the automatic rule-based assessment in case of an automatic rejection in the first place but cannot be requested right from the beginning. In case of a manual assessment, the dealer also has the possibility to send attachments (e.g., a video of rattling engine) and a free text comment along with the request.

In tangible terms, the result of the goodwill process is a percentage of the labor and parts cost contributions the dealer requests and the manufacturer is willing to pay. The set of possible contribution percentages ranges from 0 to 100% in steps of 10%: $C = \{0, 10, 20, \ldots, 100\}$. For instance, if the dealer has labor and parts costs of €1,149.82 and €903.30, respectively, and requests labor and parts cost contributions of 100%, the assessor decides which percentage of contribution is appropriate by taking all the provided information into account. He or she might first check the mileage and age of the vehicle, then the respective defect, whether the vehicle was regularly serviced, and so on. Based on these checks, he or she decides for a contribution, e.g., 50% for labor and 100% for parts. In our example, this would lead to a monetary compensation of the dealer of €574.91 for labor and €903.30 for parts.

To get an idea about the dimensions of automatic vs. manual goodwill assessments, Fig. 2 shows the overall proportion of automatic and manual goodwill assessments of some selected sales markets.

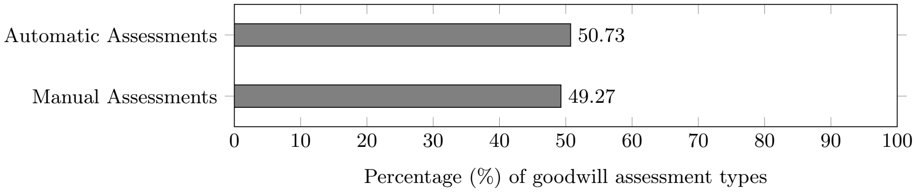

Fig. 2: Overall portions (%) of manual and automatic assessments.

Note that the period of data selection is veiled to allow no conclusions. The portion of goodwill requests that need to be assessed manually is almost as high (49.27%) as the portion of automatically processed goodwill requests (50.73%). In total numbers, 688,879 goodwill requests have been created so far, 349,488 of which were processed automatically by rules and 339,391 manually by a human expert.

Table 1 breaks down the goodwill numbers per selected National Sales Company (NSC). The NSC names have been anonymized here by letters (A to E), to prevent conclusions about goodwill strategies per country. The size of the sales market naturally influences the number of goodwill cases. From an assessment perspective it makes sense to look at the goodwill cases on a per sales market basis, since sales markets have their own goodwill strategies. Therefore, goodwill compensations is very market specific.

Table 1: Goodwill assessment numbers by National Sales Company (NSC).

| NSC | Goodwill Requests | Automatic | Manual | Degree of automation |
|---|---|---|---|---|
| A | 35,624 | 20,998 | 14,626 | 58.94 % |
| B | 76,461 | 48,666 | 27,795 | 63.65 % |
| C | 84,030 | 47,278 | 36,752 | 56.26 % |
| D | 437,656 | 200,831 | 236,825 | 45.89 % |
| E | 55,108 | 31,715 | 23,393 | 57.55 % |
| $\sum$ | 688,879 | 349,488 | 339,391 | ⌀ 50.73 % |

# 3 Prescriptive Machine Learning for Goodwill Assessment

In this section, we propose to extend the standard goodwill assessment process as outlined in the previous section, with prescriptive ML models. First, we describe how ML models could be integrated into the existing goodwill use case. Subsequently, we evaluate how well a complex human decision process such as goodwill assessment can be covered by supervised ML.

## 3.1 Enhancing the Goodwill Assessment Process

Fig. 3 shows a goodwill use case extended by ML in comparison with the classic use case outlined in Fig. 1. The *prescriptive model assessment* can either be included in the *manual assessment* process or extend the *automatic rule-based assessment*.

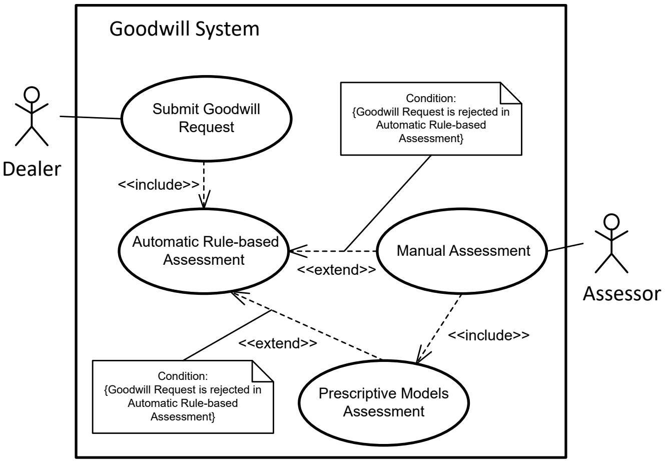
Goodwill System
Dealer
Submit Goodwill Request
<<include>>
Condition: {Goodwill Request is rejected in Automatic Rule-based Assessment}
Automatic Rule-based Assessment
<<extend>>
Manual Assessment
Assessor
<<include>>
<<extend>>
Condition: {Goodwill Request is rejected in Automatic Rule-based Assessment}
Prescriptive Models Assessment

Fig. 3: UML Use Case Diagram for the ML-enhanced goodwill process.

In the inclusion scenario, the prescriptive model supports the manual assessment through goodwill contribution suggestions that guide the assessor in his or her decision process. The prescriptive model serves as a *decision support system* (DSS) and only informs the assessor about the presumably most appropriate decision. Accepting the decision is not compulsory for the assessor, who still possesses the sovereignty over the goodwill decision. Nevertheless, the model suggestions could help to harmonize and standardize decisions from a business perspective. Including the prescriptive model assessment in the manual assessment might be a good starting point for making use of ML in the goodwill process, as the risk of wrong assessments is low and the final decision is still in the hands of an expert.

In the extension scenario, the model extends the automatic rule-based assessment and takes over cases not decidable by rules. The model assesses goodwill decisions automatically and supports the process through *automated decision making* (ADM). From a business perspective, this is the ultimate goal to aim for, as it will directly reduce process costs. However, this approach also comes with the greatest risk, as there is no human expert involved anymore who supervises the final decisions. Customer satisfaction and financial impact for the manufacturer are left to the machine. Leaving the final goodwill decision to a prescriptive model requires trust that can only be built through an evaluation by business experts over a long term period.

A combination of inclusion and extension is also conceivable. While ADM might be feasible in less complex cases, it might be advisable to just integrate the model as a DSS in more complex scenarios, leaving the final decision to a human expert. What exactly distinguishes less and more complex goodwill scenarios is still an open research question.

### 3.2 Prescriptive Machine Learning

The setting of prescriptive ML deviates from the standard setting of predictive ML in various ways [7]. This also includes the process of supervision. As already mentioned, in prescriptive ML, there is not necessarily something like a "ground-truth" or correct decision, and even if decisions might be compared in terms of quality or desirability of their implications, there is no guarantee that decisions made by human experts in the past were optimal. Therefore, taking them directly as targets for a supervised learning method might not be advisable [11]. In the case of goodwill, for example, a decision of 50% contribution appears to be somewhat overrepresented (cf. Fig. 4), letting one suspect that this is often taken as a default choice for a partial cost coverage, even if it might not necessarily be the most appropriate percentage. In the following, we will nevertheless assume that mimicking the expert is a reasonable strategy, at least as a first step toward a data-driven goodwill assessment, leaving more elaborate approaches for future work.

Under this premise, the problem is essentially reduced to a supervised learning task, with the observed human goodwill decisions

$$\mathcal{D} = \big\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\big\}$$

as training data. Instances are goodwill requests entered by the dealer and represented as a *feature vector* $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^m$. These instances are labeled by assessed contribution percentages, which serve as the target variable $y \in \mathcal{Y} \subseteq \mathbb{R}$. The goal of the ML task is to learn a decision model $h^* \in \mathcal{H}$, where $\mathcal{H}$ is the class of candidate models (referred to as hypothesis space in the common setting of supervised learning). This model is a mapping $\mathcal{X} \rightarrow \mathcal{Y}$ supposed to approximate the training data and, more importantly, generalize well to new decision problems. Like in supervised learning, we model the performance of a model $h$ in terms of a loss (error) function $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, so that $l(y, \hat{y})$ denotes the penalty incurred by the learner for prescribing $\hat{y}$ when the expert decides $y$. The choice of a presumably optimal model $h^*$ is commonly guided by the empirical risk

$$R(h) := \frac{1}{n} \sum_{i=1}^{n} l(y_i, h(\boldsymbol{x}_i)) \tag{1}$$

as an estimate of a model's performance. This measure is normally not minimized directly by the learner, however, because the empirical risk minimizer $h^* = \arg\min_{h \in \mathcal{H}} R(h)$ is knowingly prone to overfitting the training data, and hence to suboptimal generalization.

### 3.3 Human Goodwill Decision Data

Table 2 shows the features used for the ML task. In the first step, we will only look at the *hard facts*, such as vehicle mileage, vehicle age, the defect code, the costs, and the requested labor and part contributions. The raw data entered by the dealer will be enriched with further vehicle data that can be derived from the vehicle identification number (VIN), including the vehicle model type, the series, the motor series, the order country of the vehicle, the sales country of the vehicle, and whether the vehicle is a car or motorbike. The free-text dealer comment and attachments will be ignored for now, because they can be considered as “soft” facts. Besides, they are not immediately usable and require sophisticated post-processing techniques such as NLP. The rest of the data is a mixture of categorical and numerical data and qualifies as tabular data.

The features are pre-processed as follows: Numeric data is scaled using *min-max-scaling* (e.g., Parts, Labor and Total Costs), low cardinality categorical features are encoded using *one-hot-encoding* (e.g., Customer Type or Requested Labor and Parts Contributions ), and high cardinality features are *hashed* (e.g., Defect Code or Vehicle Series).

Turning our attention to the target variable, Fig. 4 shows how the overall contributions are distributed over the possible percentages $\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$. Obviously, the data is heavily imbalanced, and contributions other than 0% and 100% are rarely used. Among the rare contributions, the 50% decision sticks out and appears a bit more frequently, whereas 90% is the least frequent contribution. As already said, this may reflect a common human pattern: If not being exactly sure what to grant, people tend to opt for a compromise in the middle. Another pattern one can observe is a kind of “generous rounding” to

Table 2: Features used for model training.

| Attribute | Data Type | Description |
|---|---|---|
| Vehicle Mileage | Numeric (continuous) | 12,500 |
| Vehicle Age | Numeric (continuous) | 48 |
| Enquiry Indicator | Categorical (ordinal) | Request after or before the repair |
| Warranty Stage | Categorical (nominal) | Standard or Extended Goodwill |
| Product Type | Categorical (nominal) | Car or Motorbike |
| Regular Service | Categorical (nominal) | Yes or No |
| Sales Country | Categorical (nominal) | NL |
| Order Country | Categorical (nominal) | BE |
| External Guarantee | Categorical (nominal) | Yes or No |
| Vehicle registered to customer | Categorical (nominal) | Yes or No |
| Vehicle Model Type | Categorical (nominal) | FG81 |
| Vehicle Series | Categorical (nominal) | G21 |
| Motor Series | Categorical (nominal) | N57T |
| Mobility provided | Categorical (nominal) | Yes or No |
| Defect Code | Categorical (nominal) | 1178031500 |
| Defect Code (Main and sub group only) | Categorical (nominal) | 1178 |
| Shared last expenses | Categorical (nominal) | Yes or No |
| Customer Type | Categorical (nominal) | Regular, Transit or International |
| Requested Labor Contribution (per cent) | Categorical (nominal) | 60% |
| Requested Parts Contribution (per cent) | Categorical (nominal) | 60% |
| Dealer Labor Contribution (per cent) | Categorical (nominal) | 40% |
| Dealer Parts Contribution (per cent) | Categorical (nominal) | 40% |
| Parts Costs | Numeric (continuous) | €903.30 |
| Labor Costs | Numeric (continuous) | €1,149.82 |
| Requested Open Time Units | Numeric (discrete) | 5 |
| Dealer Open Time Units | Numeric (discrete) | 2 |
| Additional service costs, e.g., replacement car | Numeric (continuous) | €460.30 |
| Total Costs | Numeric (continuous) | €3,682.89 |

"meaningful" contributions, namely, 0%, 30%, 50%, 70%, 100%. Other contributions, such as 10% and 90%, are even more rare, probably because these are considered somewhat pedantic. In any case, the rare contributions are likely to carry important information, as they reflect subtle human instinct, and they are key to safeguard customer satisfaction. There is also an apparent tendency to contribute rather than not contribute from manufacturer's perspective, as the 100% bar is noticeably higher than the 0% bar. This is the case for labor as well as parts. However, for parts the tendency is stronger than for labor.

### 3.4 Hierarchical cost-sensitive learning

From the description of the task and the data, it becomes clear that goodwill assessment comes with a number of important challenges from a machine learning perspective. First, looking at the scale of the target variable (contribution

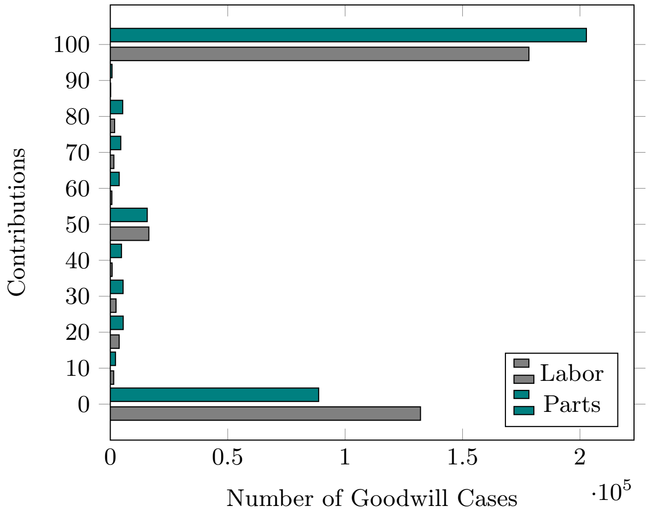

Fig. 4: Distribution of goodwill contributions for Labor and Parts at BMW.

in percentage), the problem is somehow in-between ordinal classification and regression: In principle, the target is numerical, but not all numbers between 0 and 100 are deemed valid prescriptions. Therefore, one may also think of tackling the task as a problem of ordinal classification with 11 class labels sorted in increasing order from lowest (0%) to highest (100%).

Related to the interpretation of the scale is the question of how a suitable loss function should look like. Obviously, a standard measure such as misclassification rate (0/1 loss) is inappropriate, even if the task is treated as a classification problem, because the loss function should take the linear structure of the contribution scale into account. Squared or absolute error as commonly used in regression do not appear to be perfect choices either, as one may argue that there is not only a quantitative but also a *qualitative* difference between the 0% decision, the 100% decision, and the decision of a partial contribution. This suggests a cost-sensitive approach, in which a cost (loss) function $\mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ is explicitly defined in “tabular” form. As an additional advantage, this allows for incentivising the learner in a strategic way, e.g., to constructing more customer-friendly or more cost-oriented decision models.

Another challenge is the class imbalance. Imbalanced data makes learning more difficult, and many algorithms have a tendency to compromise the accuracy of small classes in favor of bigger classes [12]. This would be especially problematic in the case of goodwill assessment, enforcing extreme decisions at the cost of partial contributions. Common approaches to deal with imbalanced data include up-sampling of the minority classes or down-sampling of the predominant classes in order to balance the data [13]. Similar effects can be achieved by adding weights to the training examples, making the underrepresented examples more important and the overrepresented less.

To tackle both problems, cost-sensitivity and imbalance, we propose a hierarchical approach with a qualitative (categorical) first layer and a quantitative second layer. In the first layer, we solve an ordinal 3-class classification (or ranking) problem, distinguishing between classes NO (no contribution, rank 1), PARTIAL (partial contribution, rank 2), and FULL (full contribution, rank 3). Obviously, this problem is more balanced, because all contributions between 10% and 90% are collected in a single class.

In the case where an instance is assigned to PARTIAL in the first layer, it is forwarded to the second layer, where the concrete percentage of contribution is determined. Thus, while an instance $\boldsymbol{x}$ is mapped to a rank $r(\boldsymbol{x}) \in \{1, 2, 3\}$ in the first layer, $\boldsymbol{x}$ is mapped to any of the numbers $\{10, 20, \ldots, 90\}$ in the second layer. The latter task can be formalized as a (constrained) regression problem.

The first problem, where an example $(\boldsymbol{x}, y)$ consists of an input vector $\boldsymbol{x} \in \mathcal{X}$ and an ordinal label $y \in \mathcal{Y} = \{1, 2, ..., K\}$ (in our case $\{\mathrm{NO}, \mathrm{PARTIAL}, \mathrm{FULL}\}$, i.e., $K = 3$), provides us with the opportunity to use the cost-sensitive ranking framework presented in [9]. This framework allows one to specify a *cost matrix* in a flexible way, which is especially convenient in our case. In fact, by utilizing a custom defined $K \times K$ cost matrix $\mathcal{C}$, we can configure the mislabeling cost according to our strategy, e.g., rather customer-friendly or more cost-oriented from manufacturer's perspective. The cost of predicting an example $(\boldsymbol{x}, y)$ as rank $k$ is given by the entry $\mathcal{C}_{y,k}$ in the cost matrix. Table 3 shows two distinct strategies for goodwill assessments. The cost matrix on the left side shows a customer-friendly strategy, where the learner is strongly penalized when prescribing NO instead of FULL ($\mathcal{C}_{3,1} = 30$). On the right side, the cost matrix implements a more cost-orientated approach, where the learner is penalized the most for the decision FULL instead of NO ($\mathcal{C}_{1,3} = 30$). Note that the result of the regression model for the PARTIAL values ($k = 2$) will be mapped back to the interval $\mathcal{C}_{2,2} = [0, 5]$ to also integrate the regression into the overall cost-sensitive ranking framework. By the width of the interval, we can configure how much importance we give to the exact prediction of the values of the regression layer. Fig. 5 visualizes the structure of the proposed hierarchical approach.

Table 3: Different assessment strategies specified by different cost functions: customer-oriented with higher penalization of contributions that are loo low (left) vs. manufacturer-oriented with higher penalization of contributions that are too high (right).

| | | *Prescribed* | | |
|---|---|---|---|---|
| | | NO | PARTIAL | FULL |
| *Actual* | NO | 0 | 5 | 10 |
| | PARTIAL | 10 | [0,5] | 5 |
| | FULL | 30 | 10 | 0 |

| | | *Prescribed* | | |
|---|---|---|---|---|
| | | NO | PARTIAL | FULL |
| *Actual* | NO | 0 | 10 | 30 |
| | PARTIAL | 5 | [0,5] | 10 |
| | FULL | 10 | 5 | 0 |

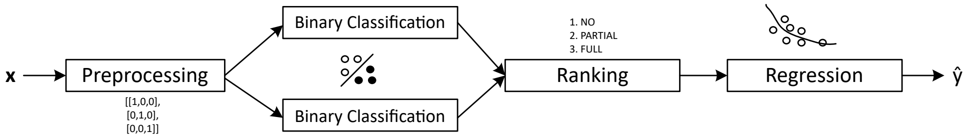

Fig. 5: Overview of the hierarchical cost-sensitive approach.

The approach [9] to ordinal classification is based on a reduction to weighted binary classification. More specifically, a binary classifier

$$f : \mathcal{X} \times \{1, \ldots, K-1\} \to \{0, 1\}$$

is trained that accepts *extended* instances $(\boldsymbol{x}, k)$ as input. As output, the classifier is supposed to produce 1 (answer "yes") if the true rank of $\boldsymbol{x}$ exceeds $k$ and 0 (answer "no") otherwise. The actual rank of a query instance can then be determined by applying the following ranking rule:

$$r(\boldsymbol{x}) = 1 + \sum_{k=1}^{K-1} f(\boldsymbol{x}, k) \,. \tag{2}$$

To train the classifier, the original data is extended as follows: Each original example $(\boldsymbol{x}, y)$ is turned into extended examples $(\boldsymbol{x}^k, y^k)$ with weights $w_{y,k}$, where[3]

$$\boldsymbol{x}^k = (\boldsymbol{x}, k), \quad y^k = [\![ k < y ]\!], \quad w_{y,k} = |\mathcal{C}_{y,k} - \mathcal{C}_{y,k+1}| \,.$$

The weights $w_{y,k}$ control the importance of an example during the training phase of the binary classifier. The higher the cost difference between two adjacent ranks, the larger the weights and therefore the importance of a particular example.

Incorporating domain knowledge, we propose the following small modification of the ranking rule (2): As the proposed contribution essentially never exceeds the contribution $q$ requested for $\boldsymbol{x}$, we set

$$r(\boldsymbol{x}) = \min \big\{ 1 + f(\boldsymbol{x}, 1) + f(\boldsymbol{x}, 2), q \big\} \,. \tag{3}$$

For the second layer of our model, any regression method can in principle be used. For the exact inference of the partial contribution values, we round and constrain the regression model's output to the set of possible contributions $\{10, \ldots, 90\}$. Also, like for the prescription of ranks, we make sure that the prescription does not exceed the requested contribution $q$:

$$\hat{y} = \min \left\{ \lfloor \frac{f(\boldsymbol{x})}{10} \rceil \cdot 10, q \right\} \tag{4}$$

[3] $[\![\cdot]\!]$ denotes the indicator function returning 1 if the argument is true and 0 otherwise.

## 4 Evaluation and Results

In this section, we evaluate our hierarchical cost-sensitive approach on BMW's goodwill data sets. For training the classifier $f$ (and ranker $r$) in the first layer, a learning algorithm is needed that is able to handle weighted examples. In our experimental study, we used extreme gradient boosting (XGBoost) [3], a versatile method that proved to work very well on tabular data and also outperforms deep neural networks in this context [10]. Another advantage is that XGBoost can be used for both classification and regression, hence we could use it for training the first as well as the second layer of our model.

Tables 4 and 5 show the results of a ten-fold cross validation in terms of the mean and standard deviation of various performance metrics. The first metric of interest is the cost of the model's prescriptions according to the underlying cost function — here, we present results for the cost matrix (a) in Table 3 (those for matrix (b) look very similar). The middle part of the matrix, i.e., the cost for assessments involving a partial contribution, is filled with the absolute error of the regression model scaled to the specified interval (in this case $[0, 5]$). As the cost values are measured on an abstract scale without interpretable dimension, we also report the mean accuracy (ACC) for the ranking part and the mean absolute error (MAE) for the regression model (on a scale from 10 to 90), thereby making the results more tangible. Overall, our model shows a quite satisfactory performance.

Table 4: Evaluation metric results obtained for Labor.

| | Ranking | | Regression | | Costs | |
|---|---|---|---|---|---|---|
| NSC | ACC | SD | MAE | SD | C | SD |
| A | 0.887 | 0.032 | 0.942 | 0.24 | 1.133 | 0.303 |
| B | 0.904 | 0.014 | 5.094 | 0.524 | 1.018 | 0.221 |
| C | 0.926 | 0.028 | 4.519 | 0.454 | 0.725 | 0.271 |
| D | 0.857 | 0.009 | 1.306 | 0.19 | 1.321 | 0.09 |
| E | 0.881 | 0.047 | 7.161 | 1.755 | 1.064 | 0.398 |
| Mean | **0.891** | 0.026 | **3.8044** | 0.6326 | **1.0522** | 0.2566 |
| Median | **0.887** | 0.028 | **4.519** | 0.454 | **1.064** | 0.271 |

As already explained, the cost function can be used to tailor a decision model to certain strategies, e.g., making it more customer-friendly or more manufacturer-friendly (cost-oriented). To evaluate this feature, we looked at the confusion matrices obtained for the cost functions in Table 3. As can be seen in Table 6, the confusion matrix for the customer-friendly cost matrix is indeed more geared to the right, showing a tendency toward higher ranks and consequently higher contributions. In contrast, the matrix for the cost-oriented

Table 5: Evaluation metric results obtained for Parts.

| | Ranking | | Regression | | Costs | |
|---|---|---|---|---|---|---|
| NSC | ACC | SD | MAE | SD | C | SD |
| A | 0.889 | 0.035 | 1.265 | 0.249 | 1.059 | 0.452 |
| B | 0.869 | 0.016 | 5.691 | 0.485 | 1.215 | 0.158 |
| C | 0.949 | 0.023 | 6.522 | 0.711 | 0.552 | 0.183 |
| D | 0.872 | 0.011 | 4.625 | 0.313 | 1.154 | 0.078 |
| E | 0.887 | 0.055 | 7.041 | 1.732 | 1.001 | 0.51 |
| Mean | **0.8932** | 0.028 | **5.0288** | 0.698 | **0.9962** | 0.2762 |
| Median | **0.887** | 0.023 | **5.691** | 0.485 | **1.059** | 0.183 |

strategy is more geared towards the left side, with lower ranks and thus less contributions.

Table 6: Different parts ranking confusion matrix depending on the assessment strategy (for NSC A): customer-oriented (left) vs. manufacturer-oriented (right).

| | | *Prescribed* | | |
|---|---|---|---|---|
| | | NO | PARTIAL | FULL |
| *Actual* | NO | 494 | 47 | 45 |
| | PARTIAL | 0 | 286 | 34 |
| | FULL | 2 | 13 | 541 |

| | | *Prescribed* | | |
|---|---|---|---|---|
| | | NO | PARTIAL | FULL |
| *Actual* | NO | 526 | 40 | 20 |
| | PARTIAL | 6 | 295 | 19 |
| | FULL | 11 | 34 | 511 |

## 5 Conclusion and Future Work

In this paper, we described the existing rule-based and manual goodwill assessment process at BMW and how it can be extended through prescriptive machine learning models. This can either happen in the form of a decision support system, automated decision making, or a combination of both. Furthermore, we proposed a hierarchical, cost-sensitive approach for learning prescriptive models from human goodwill decisions, which accounts for the specific structure of the decision space, counteracts class imbalance, and allows for tailoring strategies to different value systems and market situations (e.g., customer friendly vs. cost oriented).

Motivated by our encouraging results, we plan to address the following challenges in future work.

- *Trust and Explanation*: We noticed that business experts do not immediately trust a prescriptive ML solution. Therefore, involving business experts in the development and evaluation process is important, not only to improve the ML solution itself, but also to foster trust in it. Explainability will play a key role in this regard, making machine learning more transparent and accessible to all stakeholders involved [5]. In fact, decisions need to be explained, and different parties may have different needs for explanation. For a dealer, feedback about the most important attribute that led to the rejection of the request might be enough, whereas an auditor needs to understand the whole reasoning process in detail.
- *Uncertainty*: Although the decision models we trained perform very well, showing the high potential of automated decision making, not all decisions appear to be perfect all the time. Therefore, it would be desirable to increase the uncertainty-awareness of decision models, so that final decisions could be transferred to the human expert in cases of high uncertainty [6].
- *Weak supervision*: As already mentioned, human goodwill decisions might be biased in one way or the other and should not necessarily be taken as a gold standard. Additionally, the data may contain concept drift due to strategy changes in the assessment process over time. Therefore, past decisions should be considered and modeled as *weak* information about the target rather than an incontestable ground truth, suggesting the use of methods for weakly supervised learning [14] in prescriptive modeling.
- *Fairness:* Another important question concerns the notion of fairness in the goodwill decision process. There might be different strategies toward fairness, depending on the sales market. For instance, some markets might want to treat all customers equally, independently of the money they spent for a vehicle, whereas others might want to prefer customers with higher priced vehicles in the goodwill process. It needs to be investigated whether or not models can be tailored to such strategies automatically, or if a manual intervention is required [4].

## References

1. Abu-Naser, S.S., Bastami, B.G.: A proposed rule based system for breasts cancer diagnosis. World Wide Journal of Multidisciplinary Research and Development **2**(5), 27–33 (2016)
2. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research **70**, 245–317 (Jan 2021)
3. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016)
4. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*). p. 329–338. Association for Computing Machinery, New York, NY, USA (2019)

5. Hong, S.R., Hullman, J., Bertini, E.: Human factors in model interpretability: Industry practices, challenges, and needs. Proc. ACM Hum.-Comput. Interact. **4**(CSCW1) (2020)
6. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning **110**(3), 457–506 (2021). https://doi.org/10.1007/s10994-021-05946-3
7. Hüllermeier, E.: Prescriptive machine learning for automated decision making: Challenges and opportunities. CoRR **abs/2112.08268** (2021), https://arxiv.org/abs/2112.08268
8. Karthikeyan, R., Geetha, P., Ramaraj, E.: Rule based system for better prediction of diabetes. In: 3rd International Conference on Computing and Communications Technologies (ICCCT). pp. 195–203 (2019)
9. Li, L., Lin, H.T.: Ordinal regression by extended binary classification. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems. vol. 19. MIT Press (2006)
10. Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. Information Fusion **81**, 84–90 (2022)
11. Swaminathan, A., Joachims, T.: Counterfactual risk minimization: Learning from logged bandit feedback. In: Proc. ICML, Int. Conf. on Machine Learning. pp. 814–823 (2015)
12. Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Proceedings ICML, 24th International Conference on Machine Learning. pp. 935–942. New York, NY, USA (2007)
13. Zhang, N.N., Ye, S.Z., Chien, T.Y.: Imbalanced data classification based on hybrid methods. In: Proceedings ICBDR, 2nd International Conference on Big Data Research. p. 16–20. Association for Computing Machinery, New York, NY, USA (2018)
14. Zhou, Z.: A brief introduction to weakly supervised learning. National Science Review **5**, 44–53 (2018)