A machine learning based approach on employee attrition prediction with an emphasize on predicting leaving reasons

Fabian Engl
Vitesco Technolgies GmbH
People Analytics and Technology;
Ostbayerische Technische
Hochschule Regensburg
Siemensstraße 10-12, 93055 Regensburg

Email: fabian.engl@oth-regensburg.de

Frank Herrmann
Ostbayerische Technische
Hochschule Regensburg
Labor Wirtschaftsinformatik, SAP
und Produktionslogistik
Galgenbergstraße 32, 93053 Regensburg

Email: frank.herrmann@oth-regensburg.de

***Abstract*—Using Vitesco Technologies as an example, this article examines whether machine learning models are suitable for detecting employee attrition at an early stage, with the aim of uncovering underlying reasons for leaving. Nine different machine learning algorithms were examined: K-nearest-neighbors, Naive Bayes, logistic regression, a support vector machine, a neural network, a random forest, adaptive boosting, and two gradient boosting models. A three-way-holdout validation method was implemented to assess the quality of the results and measure both the f-score and the degree of model generalization. Initially, it was found that tree-based methods are best suited for classifying employees. A multiclass classification approach showed that under certain conditions it is even possible to predict the underlying leaving reasons.**

I. Introduction

According to a 2018 study conducted by Bund Verlag, 71 percent of German employees reported a lack of joy in their daily work, with one in seven actively considering the prospect of quitting their jobs (Bund-Verlag 2018). However, these circumstances are not limited to just one country, but rather depict a global issue that notably impacts the younger generation (Kelly 2023). This development, combined with a labor market in favor of employees, leads to an increased number of employee-sided resignations. In business, this problem is referred to as employee attrition.

This poses a range of challenges for companies such as Vitesco Technologies, with disengaged employees resulting in reduced productivity, a higher number of workplace accidents, and in the case of resignations even burnout among the remaining colleagues as they have to leverage the additional workload (Wallace 2023). Adding to that, resignations come with substantial additional costs, as replacing a position can often cost three to four times the position's annual salary (Navarra 2022). Therefore, it should be of interest for HR management to identify and address existing dissatisfaction on the side of employees and to mitigate these issues. Yet, HR decisions often rely on subjective judgments, which cannot fully encompass the complexity of these conflicts. As a result, companies are increasingly implementing data-driven approaches like machine learning (Chugani 2023).

II. Problem Description

Prior to this publication, Vitesco Technologies' did not use machine learning models in an HR context. The goal of this research is to examine whether such models can identify employees at risk of leaving and uncover associated attrition factors. This paper analyses a data set with a total of 1,500 employees from the Chinese locations Tianjin, Wuhu, Changchun and Shanghai and is structured into two parts:

First, we review machine learning algorithms that are commonly used for employee attrition prediction and identify the optimal model for our data set. Here, the emphasis is placed on relevant pre-processing steps and metrics that ensure qualitative machine learning predictions. Building on these findings, the second part of this study delves deeper into distinguishing between reasons for employee attrition. The goal is to answer whether any correlations exist within the data that can provide insights into the employees motivations for leaving.

III. Definition Employee Attrition

There exist many definitions for employee attrition, which all share a common understanding that it refers to an employee leaving without any action or influence from the company (Arqawi et al. 2022). On one hand, possible reasons for such resignations include a fundamental dissatisfaction (Srivastava and Eachempati 2021). On the other hand, retirement and premature death are also considered employee attrition (Jain and Nayyar 2018). A few publications also list internal job changes and promotions, presuppose the elimination of the position, or even account for layoffs due to poor performance or internal restructuring (Raza et al. 2022; Gim and Im 2023; Alduayj and Rajpoot 2018; Alao and Adeyemo 2013). The

subsequent prediction of employee attrition aims to identify these departures and uncover the accompanying influencing factors.

This study only considers employees who voluntarily left the company. Reasons for leaving include professional and educational development, higher salaries, a relocation (for personal reasons), problems with managers, leadership or company values, and a lack of work-life balance.

IV. Literature Review

The specific focus of employee attrition prediction studies varies depending on the author's research background. Mainly, two perspectives can be distinguished: a business-oriented perspective and an IT perspective. Business-driven publications first develop theses about the reasons for employee attrition and represent them in the form of statistically evaluable features. Here machine learning methods serve only to verify their assumptions (Srivastava and Eachempati 2021). Moreover, they usually include a high proportion of descriptive statistics from which they derive their hypotheses. Some studies also present solutions completely without the use of machine learning (Guerranti and Dimitri 2023; Pawar, Saraf and Pradhan 2023). In the IT-oriented literature, the focus lies on the implementation of machine learning models. In these articles, there is often no formulation of such hypotheses. Since this article examines employee attrition prediction from a computer science perspective, mainly economically inspired publications are not further considered in the following literature review.

Despite numerous publications, there is hardly any practice-related literature on employee attrition prediction. Due to the lack of real industry data, the vast majority of authors resort to the "IBM HR Analytics Employee Attrition & Performance" data set (Cf. et al Fallucchi et al. 2020; Alduayj and Rajpoot 2018; Najafi-Zangeneh et al. 2021; Jain and Nayyar 2018; Bhatta et al. 2022). It contains a partial extract of IBM's employee data. At the time of this article, to our knowledge, only four publications use real company data. Of these, Alao and Adeyemo, for example, analyzed attrition in Southwest Nigerian government institutions, covering a period of over 30 years. However, the data set includes only 309 resignations among 4326 entries; in contrast to roughly 750 in our data set (Alao and Adeyemo 2013). Therefore, the results of Alao and Adeyemo cannot be applied to our use case as lages companies such as Vitesco Technologies tend to have a much higher employee turnover rate.

Two other publications by Sikaroudi et al. and Srivastava and Eachempati examine employee attrition within the automotive supplier Arak and a "mid-sized fast-moving consumer goods (FMCG) company" (Sikaroudi, Ghousi and Sikaroudi 2015; Srivastava and Eachempati 2021). Both review multiple machine learning models.

However, the article on Arak only lists the accuracy of the models, which alone is not a sufficient machine learning metric. Other authors supplement it with additional metrics such as precision, recall, or the f-score (Najafi-Zangeneh et al. 2021). Adding to that the accuracy of just under 90 percent can hardly be put into comparison as no information about the size and structure of the data is given. Basic characteristics used in our approach, such as gender or salary, are also missing. Sikaroudi et al. recommend the use of Naive Bayes, which requires statistically independent characteristics (Ertel 2016).

Srivastava and Eachempati devote a considerable amount of their paper to the optimization of the algorithms hyperparameter and show how they can significantly increase the quality of predictions. The authors advise using a deep neural network because in their case it was able to correctly classify 92 percent of employees. However, they only covers seven features, including the number of projects assigned and the amount of time invested in them. It can be concluded from this that the analyzed employees work predominantly in a project-oriented manner. This way of working and the selected characteristic also do not match our present use case.

The last study based on real company data examines employee attrition of insurance agents (Valle and Ruz 2015). Yet, their intention contradicts our research goal as firstly, the authors do not consider financial or sociological employee-related factors and make their predictions purely on the basis of performance indicators. Secondly, the results also serve to identify employees who should be terminated due to poor performance, which the researchers justify with the "practical interest to the call center" (Valle and Ruz 2015). These two assumptions severely limit the transfer of the findings to other, less competitive business sectors, especially because other publications have already demonstrated clearly that correlations exist between personal well-being and termination behavior (cf. et al. Fallucchi et al. 2020; Jain and Nayyar 2018; Alduayj and Rajpoot 2018; Aggarwal et al. 2022).

Three other research groups used information from publicly available databases as well as professional social networks (Dahan et al. 2020; Alaskar, Crane and Alduailij 2019; Kisaog 2014). Here, however, quantitative as well as qualitative shortcomings severely limited the explanatory power of the models (Dahan et al. 2020; Kisaog 2014).

The remaining publications based on the IBM data differ in their structure, in some cases considerably. While researchers from real companies in particular pay little attention to data pre-processing steps such as feature engineering, they play a central role in many of the IBM publications (Cf. et al. Alduayj and Rajpoot 2018; Alao and Adeyemo 2013; Gim and Im 2023; Pawar, Saraf and Pradhan 2023; Raza et al. 2022). This lack of focus is also criticized by Najafi-Zangeneh et al., who emphasize that especially the feature selection process must fit the characteristics of the data set (Najafi-Zangeneh et al. 2021). Several of the authors conclude that careless use of features increases model complexity, complicates evaluations, and may even negatively affect validity (Alduayj and Rajpoot 2018; Alao and Adeyemo 2013). Various proposed solutions are given for feature reduction and selection approaches. These include statistical analysis or machine-learning-based approaches (Gim and Im 2023; Gopinath and Subhashini 2020; Gim and Im 2023). In addition to feature engineering, many articles initially analyze the data using visual representations

to strengthen the general understanding of the use case and to detect inconsistencies (Cf. Raza et al. 2022; Gopinath and Subhashini 2020; Najafi-Zangeneh et al. 2021).

Some authors also criticize weaknesses of the IBM data set, which has a strong imbalance between the two employee classes (Soner et al. 2022; Raza et al. 2022). This can lead to biased machine learning models (Gim and Im 2023). They solve this problem by using synthetic data to scale up to the minority class (Raza et al. 2022; Soner et al. 2022). However, they ignore the risks of this synthetic data and only discuss its benefits. According to them, for example, a large training data set has a positive influence on the quality of the results (Raza et al. 2022).

Despite numerous publications on the use of machine learning models for employee attrition prediction, research gaps remain. Most notably, there is a lack of representative contributions from companies supplying real world data. This circumstance makes it difficult to transfer existing findings to companies like Vitesco Technologies. This is mostly due to the fact that company researchers do not provide information about the knowledge base used as well as its structure. In addition, many publications neglect the relevance of a clean data basis and detailed feature engineering. Recommendations about suited machine learning models also vary so widely, that even the authors using IBM data are in disagreement. This shows that there is no universally applicably model but rather that it is essential to implement and compare multiple algorithms to identify the one most suited for the individual use case.

V. Model Validation

Before comparing the machine learning models, it must first be clarified how their suitability and the quality of their predictions can be measured. Since no reference projects exist within the HR department of Vitesco Technologies, this article first explores common validation methods and evaluates them in the context of the present use case. The f-score, which is regarded in the literature as a decisive indicator for the quality of a model, serves as the central indicator (cf. et al. Alduayj and Rajpoot 2018; Soner et al. 2022; Kisaog 2014; Bhatta et al. 2022; Raza et al. 2022). However, especially for small data sets with many features, overfitting may occur despite a high f-score (Vabalas et al. 2019). A reliance on key figures alone does not guarantee generalized model predictions. Generalization ensures that trained models can apply learned patterns and relationships to unseen data entries. To assess both generalization and the f-score, a validation process tailored to Vitesco Technologies' human resources management was developed:

Usually, holdout or cross-validation is used to validate machine learning models. Holdout validation splits the data into a training and a validation data set (Jung 2022). The model receives the former to learn the data patterns and can use the latter to verify that it has interpreted the relationships correctly. The holdout validation is well suited for measuring the degree of generalization, as the holdout data set remains unseen by the model until verification. However, the subdivision greatly reduces the size of the training data (Raschka 2020). Adding to that, it only allows for one validation cycle, because reusing the data can result in the model merely memorize the labels and thus acting in a biased manner (Raschka 2020). This lack of multilevel validation as well as the reduced training base, can lead to large variations in the results which a reliable model evaluation.

Cross-validation, on the other hand, allows for multiple training cycles using the same data set. To do this, it first divides the data set into k subsets, with each one once serving as the validation data set for the remaining training data. In the literature, a typical value for k lies between 5 and 10, since too low values also lead to the problems mentioned before (Raschka 2020). In this way, cross-validation trains k different models, which are then merged. The averaging counteracts strong fluctuations in key figures like the f-score. This is particularly beneficial for small data sets, as more entries remail for model training (Vabalas et al. 2019). In practice, cross-validation is often complemented by stratification, which ensures that each partial data set represents all existing classes equally (Berrar 2018). Models trained using cross-validation usually exhibit less bias compared to holdout validation (Raschka 2020). However, multiple uses of the data can lead to overfitting, especially when class entries are very similar (Varma and Simon 2006).

None of the presented validation approaches are able to simultaneously ensure generalization and prevent overfitting when the data sample size is small. A combination of both approaches, on the other hand, minimizes their drawbacks: A Three-Way-Validation-Method makes more optimal use of the training data while ensuring at the same time measuring the model generalization (Raschka 2020). It is based on an initial (stratified) data split followed by a cross-validation of the training data. The entire validation process is depicted in figure 1.

VI. Data Pre-Processing

The quality of machine learning predictions also strongly depends on the available information and its quality (Batista and Monard 2002). Accurate predictions require a clean as well as conclusive data set. This severely limits the use of unprocessed system extracts or raw data, which in practice usually has errors, gaps or simply insufficient expressiveness (Keim et al. 2006). Given this context, it is crucial to conduct an initial quality assessment to identify any content deficiencies or gaps before proceeding with the implementation.

A. Visual Data Exploration

Visual data exploration (short VDE) is a graphical analysis process that provides deeper insights into the data structure and allows for initial conclusions about structural discrepancies (Keim 2001). It is particullary suited as an initial tool to reveal existing correlations and provide a general understanding of

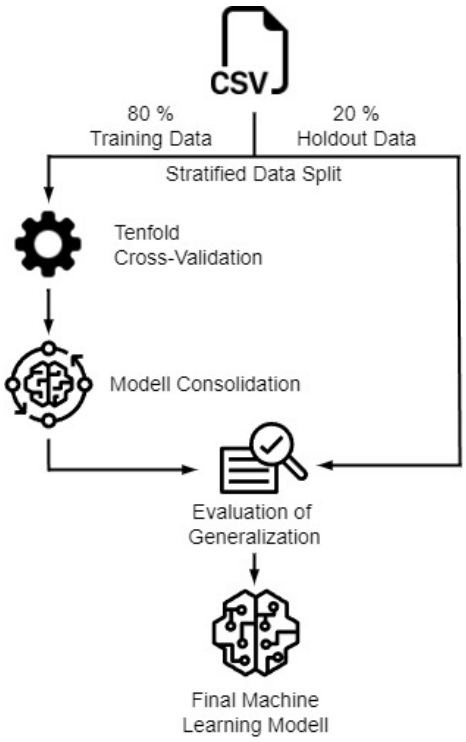

Fig. 1. Three-Way Holdout Validation Process

the use case (Cox 2017). The VDE can uncover irregularities and errors in the database, in particular, contradictory statistical distributions and outliers (Gabler 2008; Datta and Davim 2022). Highlighting and correcting these is important because undetected erroneous entries can critically affect the final machine learning predictions (Cox 2017).

In the case of Vitesco Technologies, it primarily uncovered errors of the manual data preparation such as duplicate identifiers, incorrect categories, and unrealistic entries such as a contract end date in the year 2121. Such irregularities are common as real-world industry data, to some degree, usually contains erroneous or out-of-date data (Keim et al. 2006). Nevertheless, the VDE alone is not sufficient enough to eliminate all structural deficiencies, as especially gaps in the data remain. Such blanks and NULL-values were filled using a KNN imputation. This approach is widely regarded as the most suitable and versatile method for data imputation (Emmanuel et al. 2021; Ismail, Abidin and Maen 2022). While other algorithms, particularly regression methods, may yield slightly more precise results, they often demonstrate only marginal improvements in direct comparison (Makaba and Dogo 2019). Hence, we decided to forgo a comparison of multiple approaches.

B. Feature Engineering

Classification models try to divide data entries into different classes based on their features. For structured, relational data such as in the present use case, each column – except for the class label – represents a feature. For effective model training, these must be both easily machine-processable and -interpretable (Jung 2022). Feature engineering significantly contributes to increasing the quality of information by reducing the (raw) data to only the most relevant information. For this reason, many authors refer to it as an essential part of data pre-processing (Verdonck et al. 2021; Duboue 2020; Najafi-Zangeneh et al. 2021). The process of feature engineering consists of two parts: First, it requires the examining and transforming of the data into meaningful features to best represent the use case (Fallucchi et al. 2020). Secondly, each of these columns is analyzed to determine its relevance and is removed in case it's not relevant for the classification (Alao and Adeyemo 2013). In summary, feature engineering translates a use case from the real world into a knowledge base that can be interpreted by machines and thus acts as a link between the business view and the data-driven machine learning models (Duboue 2020).

In HR especially, decisions are based on subjective assessments of employees (Fallucchi et al. 2020). This introduces bias that can harm the correct prediction of employee attrition. Here subjective opinions may cause existing patterns to be ignored because they contradict human expectations. For example, Jain and Nayyar uncovered that employeess in their use case were more likely to quit the closer they lived to their work location, and Arqawi et al. noted that in their data, there was no correlation between employee performance and resignations (Jain and Nayyar 2018; Arqawi et al. 2022). Both results are contrary to conventional leaving reasons. Feature engineering, on the other hand, ensures an objective approach and combines both perspectives by first representing business assumptions in the form of data and then having a machine learning algorithm identifying existing patterns. A clear data set containing representative features also helps HR employees to better interpret the final results of machine learning models (Duboue 2020). Interpretability plays a significant role as employee attrition can only be counteracted if dissatisfaction factors are identified at an early stage and appropriate measures are derived from them (Fallucchi et al. 2020). For this reason, internal HR experts and their know-how were integrated into our feature engineering process to better map reasons for termination to available system information and to elaborate new features (For questions about the data set and the developed features, please contact Fabian Engl using the contact details provided).

After the feature engineering and creation phase, the relevance of the features must be evaluated (Alao and Adeyemo 2013). In some circumstances, the presence of features actually harms the performance of the models. This is particularly relevant in the case of Vitesco Technologies as several models are going to be implemented and the optimal underlying data bases differ depending on the algorithm (Duboue 2020). This evaluation can be done in three different ways: using filters, wrappers, or embedded in the training phase of the machine learning models (Gim and Im 2023). In the case of Vitesco Technologies, a backwards feature selection approach was used followed by an ordinal data encoding.

VII. Binary Classification

The machine learning models were implemented using the Python libraries scikit-learn and XGBoost, with random states guaranteeing a constant evaluation, with the latter ensuring deterministic results and thus enabling evidence-based comparison (scikit learn 2023).

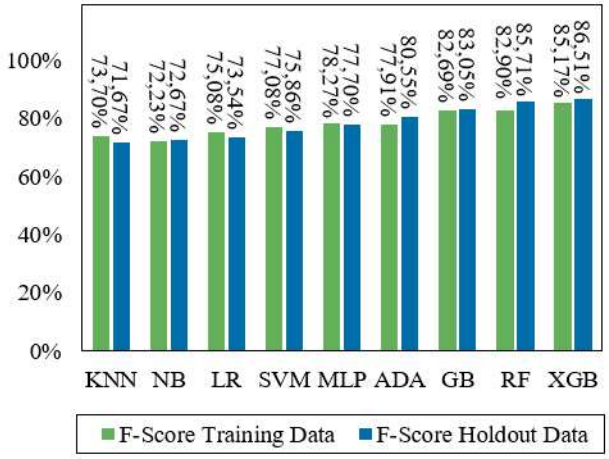

Fig. 2. F-scores of the binary machine learning models after feature engineering

A first comparison using un-optimized model shows that all achieve similar results in both cross- and holdout validation, which indicates that the holdout data represents the training data well. All machine learning models detected patterns in the HR data and were able to classify unseen data sets using the derived patterns. Overall, the tree-based algorithms – the Random Forest (RF), adaptive boosting (ADA) and both gradient boosting methods (GB and XGB) – lead the comparison. All four models achieve f-scores above 80 percent on the holdout data set (see Figure 3). However, the degree of generalization differs.

The K-nearest-neighbors (KNN) algorithm, the Naive Bayes (NB), and logistic regression (LR) record below-average results in direct comparison. Their f-scores mostly lie below 75 percent. In the case of the KNN approach, these scores result from the spatial distribution of the employee data within the feature space. Using a principle component analysis to display the feature space in a two-dimensional scatter plot, large-scale overlaps of both employee classes become apparent (see Figure 3). The entries are so close to each other that an evaluation based on neighbors does not allow for meaningful conclusions. Since the KNN algorithm, bases its predictions solely on this spatial distance of the features (Ertel 2016), it is unsuitable for the current use case.

The Naive Bayes model also underperforms in a direct comparison, which was to be exspected as the model presupposes the independence of all features (Ertel 2016). Especially in the context of employee data, this is rarely the case. For example, in Vitesco Technologies' HR systems, certain characteristics such as compensation are calculated from a combination of other information such as the number of service years. In

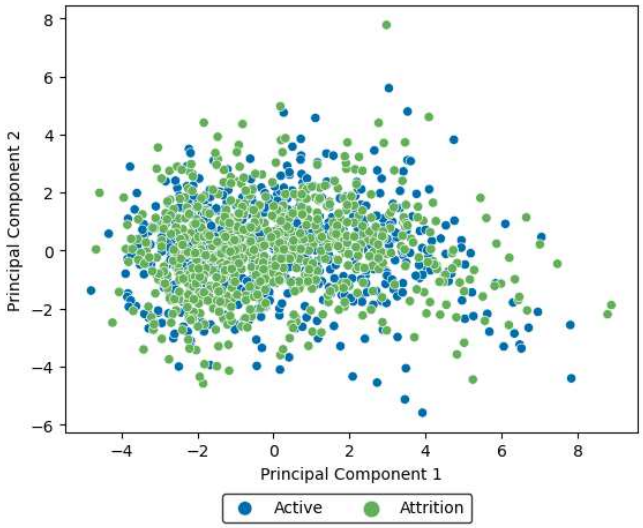

Fig. 3. Two-dimensional visualization of the KNN feature space using a principle component analysis

theory, the model is generally not suitable as a tool for employee turnover forecasting. Nevertheless, some researchers achieved accurate forecasts using Naive Bayes (Valle and Ruz 2015). In our use case, the underwhelming results and doubts about the model's suitability are enough to rule it out as unsuitable.

The logistic regression also only achieved f-scores of 75.08 and 73.54 percent. This is because the data set has many binary-categorical columns, which make it difficult to calculate clear boundaries within features (Kemalbay and Korkmazoğlu 2014). Similar to the KNN algorithm, the logistic regression seems incompatible with the employee data studied.

The support vector machine (SVM) achieves f-scores of 77.08 and 75.86 percent but shows a rather high difference between training and holdout data. However, algorithms – especially the SVM – have hyperparameters that allow for optimization of the predictions (see chapter Hyperparameter Optimization). In case of the SVM these may allow the construction of more representative hyper-planes possibly increasing the f-scores further.

A similar trend can be observed for the neural network (MLP), which comes close to matching the performance of tree-based models achieving f-scores of 78.27 and 77.70 percent, respectively. With a difference of 0.57 percent, it also records a high degree of generalization. However, there is still a significant performance gap compared to the leading machine learning model.

Since both AdaBoost and Random Forest achieve significantly higher f-scores in the holdout data set, there is reason to believe that the holdout data contains more clearly distinguishable employees and therefore slightly positively biases the results. The extreme gradient boosting procedure, which also achieved better holdout values, supports this hypothesis. Nevertheless, the comparison clearly shows that without hyperparameter tuning the tree-based methods achieve the best

results in the case of Vitesco Technologies and stand out in particular due to their high precision (For a more detailed list of all key figures as well as questions regarding the implementation, please contact Fabian Engl using the contact details provided).

A. *Hyperparameter Optimization*

Hyperparameter optimization allows to further improve the performance of machine learning methods by tuning the model to the available data and its structure (Duboue 2020). Hyperparameters are adjustment screws that make minor changes in how a model operates, thus enhancing performance (Bhatta et al. 2022). They are set before model training and do not dynamically adjust during the validation process (Nokeri 2021). The number of parameters depends on the algorithm.

The two most common methods for determining hyperparameters include a grid search and a random search (Agrawal 2021). The gird search uses a list of values for each parameter, checks all possible combinations and returns the best combination (Bhatta et al. 2022). While this approach guarantees optimal parameter matching within the given values, it requires a lot of computational resources to do so. If the data set contains many features or if a model has many parameters, this process can take several days or even weeks (Ertel 2016). The runtime of the random search, on the other hand, is unaffected by both and always remains constant (Agrawal 2021). Instead of fixed parameter values, it is given plausible value ranges for each parameter and a fixed number of iterations. It chooses random hyperparameters within these ranges and repeats this process as many times as specified. This way, it usually finds a near best set saving a huge amount of time(Agrawal 2021).

In our case limited computational capacity restricted a large-scale grid search, so both methods were combined into a two-step search: A random search first narrowed down the value ranges of the parameters as best as possible, ensuring that the resulting combination is as close to the actual optimum as possible. Afterwards, a grid search examined the immediately adjacent range to fine-tune the parameters. This approach allows for a significantly larger initial search space to be covered while reducing the drawbacks of both search methodologies.
The optimized results show that some methods benefit more from hyperparameter optimization than others (see Figure 4). While AdaBoost and the normal gradient boosting model achieved a significant performance increase, most of the other models only saw minor improvements. For the neural network, the random forest and the extreme gradient boosting method, it even lead to slightly lower holdout results. However, this development does not necessarily mean that the overall performance worsened, as the holdout data primarily serves to evaluate the generalization of the machine learning model in our validation process. Consequently, a qualitative machine learning model must both achieve a high f-score in the cross-validation as well as comparable results with the holdout data. Considering this, all models recorded better overall results, which are reflected either in an improvement of the holdout f-scores or in the degree of generalization.

The Random Forest was able to reduce the difference between the cross-validation and the holdout evaluation by 1.57 percent while both scores themselves showed only marginal improvements. The adaptive and the normal gradient boosting models showed the biggest improvements with an average improvement of 5.77 and 2.83 percent respectively. In the case of adaptive boosting a modification of tree-depth resulted in performance on part with the other leading tree-based machine learning models. This result is based on a maximum depth of three. This modification of the depth changes the basic approach of the algorithm, as it is originally based on weak learners – so-called decision stumps (Freund and Schapire 1997). Nevertheless, this adjustment leads to the highest holdout f-score and the most generalized prediction among all models. The neural network and the SVM, on the other hand, fail to catch up to the tree-based models even after hyperparameter optimization and are therefore neglected in the following chapters.

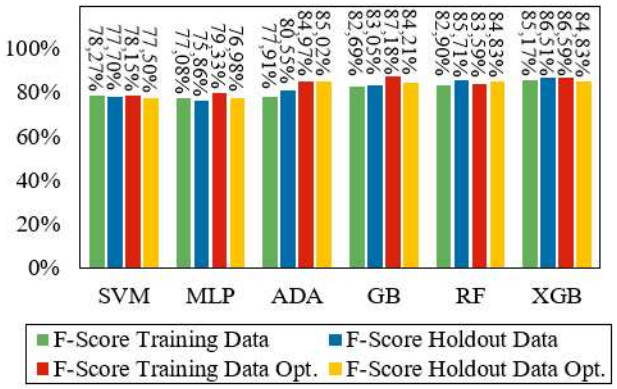

Fig. 4. F-scores of the machine learning models after the hyperparameter optimization

B. *Conclusion on binary employee attrition prediction*

After hyperparameter optimization, the tree-based machine learning models stand out as the most accurate models. It can be clearly stated that the KNN method, the Naive Bayes, the logistic regression, the neural network and the support vector machine are either incompatible or under-performing in our use case. While none of the other algorithms achieved reliable f-scores above 80 percent using our data, the results of the tree-based models stagnated at around 85 percent. Yet, due to the neglectable differences between them, no clear recommendation can be given as all seem equally suited to predict employee attrition prediction at Vitesco Technologies.

Although in our case the tree-based methods are generally suitable for employee attrition prediction, they differ in their data pre-processing and model optimizing requirements. Especially in the context of HR the choice of a machine learning model depends on more factors than just the f-score or degree of generalization. Internal resources, IT know-how and available technical resources contribute significantly to the long-term success of data-driven employee attrition detection

strategies. In this study, boosting-based models required less extensive feature engineering, which could prove beneficial if resources are limited. In particular, pre-built frameworks such as XGBoost achieve accurate predictions even with little hyperparameter optimization.

VIII. Multiclass Classification

The first binary employee attrition classification showed that machine learning models were able to correctly classify employees at risk of leaving with an 85 percent certainty. Yet this simple differentiation does not give any insights into the underlying motives for leaving which need to be revealed and understood in order to introduce HR guidelines to counteract employee attrition. For this reason this chapter wants to answer whether data patterns exist that also allow for a multi-class classification of the leaving reasons. This requires a modification to the validation process as the f-score in it's used form is only suitable for binary classification problems (Hossing and Sulaiman 2015). This chapter is uses the macro-f-score which calculates individual f-scores at class level and then calculates the average (Grandini, Bagli and Visani 2020).

The available leaving reasons are based on voluntary information provided by employees. The extend to what they reflect the reality can not be confirmed. Vitesco Technologies distinguishes between the following leaving reasons: a higher salary (SAL), better career opportunities (CAR), further education (FUE), a relocation (REL), a lack of work-life balance (WLB) and problems with the company, its culture or the leadership (CCL). The class of active employees (ACT) remains unchanged. This detailed split of former employees leads to a strong class imbalance problem as these six new classes were previously combined into one. Such strong imbalances can have a negative impact on machine learning results (Nguyen, Cooper and Kamei 2011).

A. *Imbalanced-Class-Problem*

An imbalanced-class problem exists when one or more classes contain significantly more entries than the rest (Tharwat 2020). In the case of Vitesco Technologies, this affects the class of active employees. This condition can distort machine learning metrics and hurt predictions (Najafi-Zangeneh et al. 2021). The imbalance can be counteracted by under- or oversampling the data set. Undersampling involves reducing the size of the majority class(es) to the number of entries in the smallest class, while oversampling instead increases the size of all the minority class(es) using synthetically generated data entries (Chawla et al. 2002).

B. *Initial Evaluation of patterns*

A random oversampling algorithm was used for the initial validation to analyze whether leaving-reason-specific correlations exist within the data set. Random oversampling uses existing data points to generates new entries with similar features (IBM imbalanced-learn 2023).

This initial analysis clearly shows that all models fail to differentiate all classes correctly. The f-scores regarding the class of the active employees stays mostly unchanged. While resignations due to better salary or a career advancement can still partially be identified, the remaining four reasons were not detect at all or only very unreliably. Resignations due to relocation or as a result of a lacking work-life balance were detected by the gradient boosting models, but only so inconsistently that their predictions provide little to no value for HR. With maximum f-scores of 9.52 and 19.09 percent, respectively, the results are basically on part with or just slightly better than random guessing.

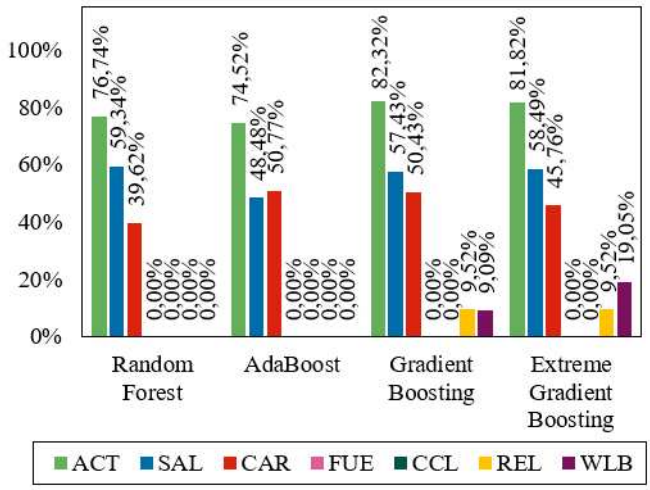

Fig. 5. Holdout f-scores of the machine learning models after applying random oversampling

The fact that only the first two leaving reasons are recognized might be due to the purely company-sided representation of the employees in the data set. All of the last four classes represent leaving reasons that are driven by intrinsic motivation. Factors such as the concordance of personal beliefs with Vitesco Technologie's company values or the experienced leadership style are based on individual perception and can vary between employees. The company-sided data set does not contain any features that represent individual well being and therefore, does not provide sufficient information to clearly identify leaving reasons affected by it. Adding to that the last four classes make up for only roughly 13 percent of all resignations. Such small classes could hurt performance even further as too few entries exist for the machine learning models, to derive clear patterns. Since both facts hinder a clear differentiation of all classes, the following multi-class classification focuses on first three classes only: active employees and former employees who left due to a higher salary or better career advancement opportunities.

C. *Comparison of under- and oversampling techniques*

Based on the selected classes, several under- and oversampling methods were compared. In direct comparison, the undersampling algorithms achieve significantly lower macro-f-scores and show higher fluctuations between the machine learning

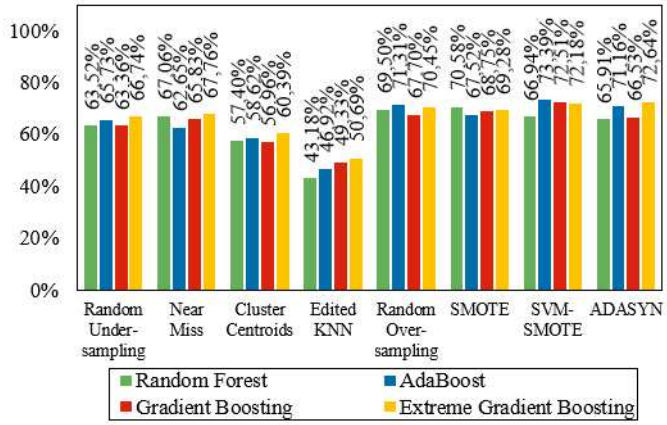

Fig. 6. Holdout f-scores of the machine learning models after applying different under- and oversampling techniques

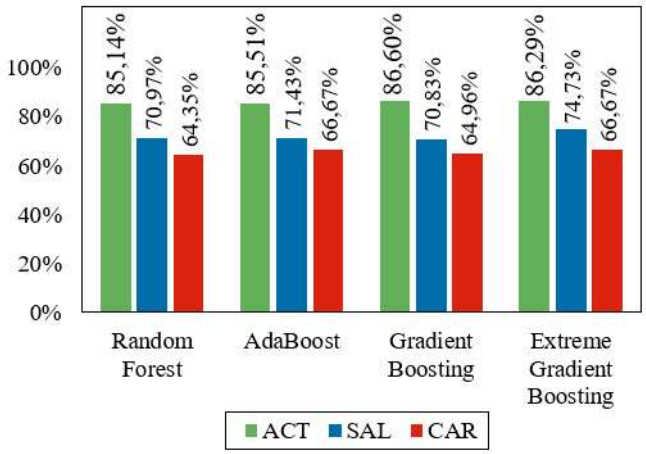

Fig. 7. Holdout f-scores of the machine learning models on class level

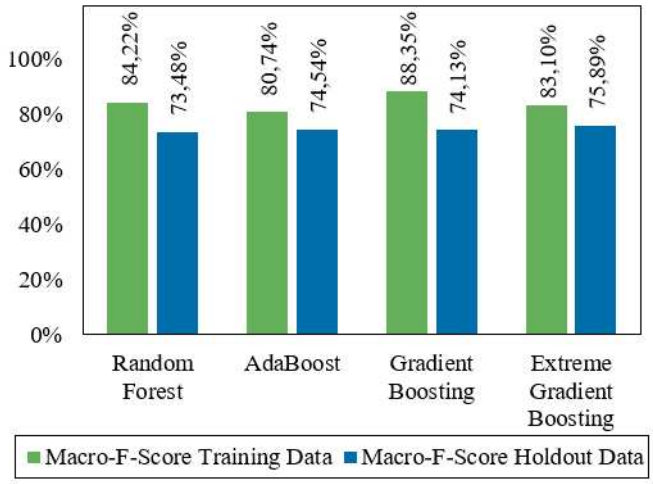

Fig. 8. Macro-f-scores of the machine learning models

models (see figure 6). Of all oversampling algorithms, the SVMSMOTE achieves the most stable macro-f-scores accross all models. For this reason it is used to counteract the class imbalance for the following comparison.

D. Evaluation of the multi-class classification

Just as in binary classification, all four machine learning models underwent both a feature engineering phase and a hyperparameter optimization. Similar to the previous chapter, the tree-based machine learning models exhibit only slight variations, with adaptive boosting achieving the highest level of generalization. The extreme gradient boosting model in particular achieves the best f-scores across all classes achieving 86.29, 74.73 and 66.67 percent. However, all models show a noticeable decline in generalization ranging from 6.21 to 14.22 percent. This probably results from the use of synthetic data.

Looking only at the active employees the class f-scores match the the binary classification results. This means they were identified with a certainty of roughly 85 to 86 percent again. This was to be expected as the class was neither modified nor affected by the excluded employee entries. Employees who resigned due to a higher salary or better career opportunities could only be detected in roughly 71 to 75 and 65 to 67 percent respectively. Although these f-scores may appear relatively low in comparison, it's crucial to consider them within the broader context of employee attrition prediction. Given that active employees are accurately classified in over 85 percent of cases, this indicates that the two primary classes of active and former employees are quite distinct. Considering this, the machine learning models can generally differentiate employees who have left Vitesco Technologies and even correctly predict the corresponding leaving reasons in more than two thirds of cases.

E. Conclusion on multi-class employee attrition prediction

In summary, machine learning models are not only capable of identifying employees at risk of leaving but also offer some insight into their leaving motives. In the case of Vitesco Technologies, the models consistently identified employees who left due to higher salary or better career opportunities. However, despite these reliable (partial) findings, all machine learning models struggle to predict all leaving reasons. One potential explanation lies in the choice of features, which may not adequately capture the employee's perspective. Many of the collected features were constructed based on a company's perspective on employee attrition and do not encompass subjective factors like the well-being of employees. Expanding the data set to incorporate additional attributes reflecting employee well-being could solve this problem. A possible solution could be asking employees for individual feedback and experiences in their exit survey.

IX. Final conclusion

In summary, the results of this paper show that machine learning models can detect distinct patterns in Vitesco Technologies' workforce data and accurately identify employees at risk of leaving. A combination of cross-validation and holdout validation is used to assess the quality of the evidence. The

(macro-)f-score is the primary metric of comparison because it most reliably reflects the predictive and differentiating capabilities of the algorithms.

Due to structural errors in the system extraction, extensive data pre-processing and preparation was necessary prior to implementation of the machine learning algorithms. Inconsistencies were first detected and corrected using a visual data exploration technique. A KNN-imputation subsequently completed structural gaps. Adding to that, the different machine learning models required varying degrees of feature engineering and hyperparameter optimization.

The comparison clearly showed that tree-based machine learning approaches are best suited for the classification of Vitesco Technologies' employee data. The random forest, the AdaBoost algorithm and both gradient boosting models achieved f-scores of around 85 percent. Therefore, all are equally suitable for employee attrition prediction. However, the extend of necessary data pre-processing steps varied significantly. In this study boosting-based approaches required less detailed feature engineering. The findings showed that the XGBoost framework achieved both a high f-score and a pronounced degree of generalization even without intensive feature engineering or hyperparameter optimizations.

In addition to identifying employees at risk of leaving, the multi-class classification approach also allowed partial conclusions to be drawn about employees motives. However, characteristics that depicted the individual well-being of the employees were missing here. As a result, it was not possible to adequately represent all reasons for leaving. Only departures due to a higher salary or better promotion prospects were reflected in the data. However, the machine learning models were ab to identify both with certainty of around 75 and 67 percent respectively.

The findings of the machine learning algorithms allow for an evidence-based alignment of Vitesco Technologies' Chinese HR policy with the employee needs. However, before the findings can be transferred to other countries and locations of Vitesco Technologies, an evaluation of the cultural influences on the results of the machine learning models is required. Furthermore, there are additional challenges to overcome: both the limited computing resources and the small data set restricted the evaluations in many places. Therefore, before integrating the proposed algorithms into the HR landscape the research results should first be validated during an initial test phase verifying the findings using real employee data and adjusting the models in case of deviations. In order to reduce manual data preparation in the future and ensure qualitative predictions, an automated data pipeline needs to be implemented.

X. Summary

Using Vitesco Technologies as an example, this article examines whether machine learning models are suitable for detecting dissatisfaction on the part of employees at an early stage and thus being able to preventively counteract terminations. Tree-based methods were found to be the most suitable for classifying employee data. These include a random forest, the AdaBoost algorithm and two different gradient boosting models. Finally, it was shown that a prediction of the associated reasons for termination is also possible under certain conditions.

References

Aggarwal, S. et al. (2022). "Employee Attrition Prediction Using Machine Learning Comparative Study". In: *Intelligent Manufacturing and Energy Sustainability*. Singapore: Springer, pp. 453–466.

Agrawal, T. (2021). "Hyperparameter Optimization Using Scikit-Learn". In: *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*. Berkeley, USA: Apress, pp. 31–51.

Alao, D. and A. B. Adeyemo (Mar. 2013). "Analyzing Employee Attrition using Decision Tree Algorithms". In: *Computing, Information Systems, Development Informatics and Allied Research Journal* 4.1, pp. 17–28.

Alaskar, L., M. Crane, and M. Alduailij (2019). "Employee Turnover Prediction Using Machine Learning". In: *Advances in Data Science, Cyber Security and IT Applications*. Cham, Schweiz: Springer International Publishing, pp. 301–316.

Alduayj, S. S. and K. Rajpoot (Nov. 2018). "Predicting Employee Attrition using Machine Learning". In: *2018 International Conference on Innovations in Information Technology*, pp. 93–98.

Arqawi, S. M. et al. (Nov. 2022). "Predicting Employee Attrition and Performance using Deep Learning". In: *Journal of Theoretical and Applied Information Technology* 100.21, pp. 6526–6536.

Batista, G. and M. C. Monard (Jan. 2002). "A Study of K-Nearest Neighbour as an Imputation Method." In: 30, pp. 2–11.

Berrar, D. (Jan. 2018). "Cross-Validation". In: *Encyclopedia of Bioinformatics and Computational Biology* 1, pp. 542–545.

Bhatta, S. et al. (2022). "Machine Learning Approach to Predicting Attrition Among Employees at Work". In: *Artificial Intelligence Trends in Systems*. Cham, Schweiz: Springer International Publishing, pp. 285–294.

Bund-Verlag (Sept. 2018). *Innere Kündigung - Fünf Millionen Arbeitnehmer haben innerlich gekündigt*. URL: https://www.bund-verlag.de/aktuelles~F%C3%BCnf-Millionen-Arbeitnehmer-haben-innerlich-gek%C3%BCndigt~.html (visited on 07/24/2023).

Chawla, N. V. et al. (June 2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Chugani, J. (July 2023). *How HR Professionals Can Coexist With AI Tools*. URL: https://www.forbes.com/sites/forbeshumanresourcescouncil/2023/07/12/how-hr-professionals-can-coexist-with-ai-tools/ (visited on 07/24/2023).

Cox, V. (2017). *Translating Statistics to Make Decisions*. 1st ed. Salisbury, UK: Springer Link, pp. 41–100.

Dahan, S. et al. (May 2020). "Predicting Employment Notice Period with Machine Learning: Promises and Limitations". In: *McGill Law Journal* 65, pp. 711–753.

Datta, S. and J. P. Davim (2022). *Machine Learning in Industry*. 1st ed. Cham, Schweiz: Springer, pp. 5–6.

Duboue, P. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*. 1st ed. Cambridge, UK: Cambridge University Press, pp. 21–43.

Emmanuel, T. et al. (Oct. 2021). "A survey on missing data in machine learning". In: *Journal of Big Data* 8, pp. 1–37.

Ertel, W. (2016). *Grundkurs Künstliche Intelligenz*. 4th ed. Wiesbaden: Springer Fachmedien, pp. 207–209, 237–239, 298, 304.

Fallucchi, F. et al. (Dec. 2020). "Predicting Employee Attrition Using Machine Learning Techniques". In: *Computers* 9.4, pp. 86, 1–17.

Freund, Y. and R. E. Schapire (Aug. 1997). "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55, pp. 119–139.

Gabler, T. (2008). "Auffälligkeiten in Häufigkeitsverteilungen". In: *Wirtschaftsstatistik im Bachelor - Grundlagen und Datenanalyse*. 1st ed. Wiesbaden: Springer Link, pp. 117–148.

Gim, S. and E. T. Im (2023). "A Study on Predicting Employee Attrition Using Machine Learning". In: *Big Data, Cloud Computing, and Data Science Engineering*. Cham, Schweiz: Springer International Publishing, pp. 55–69.

Gopinath, R. and M. Subhashini (Dec. 2020). "Employee Attrition Prediction in Industry using Machine Learning Techniques". In: *International Journal of Advanced Research in Engineering & Technology* 11, pp. 3329–3341.

Grandini, M., E. Bagli, and G. Visani (Aug. 13, 2020). *Metrics for Multi-Class Classification: an Overview*. arXiv: 2008.05756[cs,stat].

Guerranti, F. and G. M. Dimitri (Jan. 2023). "A Comparison of Machine Learning Approaches for Predicting Employee Attrition". In: *Applied Sciences* 13.1, pp. 267, 1–8.

Hossing, M. and M. N. Sulaiman (Mar. 2015). "A Review on Evaluation Metrics for Data Classification Evaluations". In: *International Journal of Data Mining & Knowledge Management Process* 5.2, pp. 1–11.

IBM imbalanced-learn (2023). *Naive random over-sampling*. URL: https://imbalanced-learn.org/stable/over_sampling.html (visited on 07/24/2023).

Ismail, A. R., N. Z. Abidin, and M. K. Maen (Feb. 2022). "Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare". In: *Journal of Robotics and Control* 3.2, pp. 143–152.

Jain, R. and A. Nayyar (Nov. 2018). "Predicting Employee Attrition using XGBoost Machine Learning Approach". In: *2018 International Conference on System Modeling & Advancement in Research Trends*, pp. 113–120.

Jung, A. (2022). *Machine Learning: The Basics*. 1st ed. Singapore: Springer Nature, pp. 3–22, 117–142, 172.

Keim, D. A. (Aug. 2001). "Visual exploration of large data sets". In: *Communications of the ACM* 44.8, pp. 38–44.

Keim, D. A. et al. (July 2006). "Challenges in Visual Data Analysis". In: *Tenth International Conference on Information Visualisation*, pp. 9–16.

Kelly, J. (Feb. 2023). *Workers In China Have Their Own Version Of Quiet Quitting And Acting Your Wage: 'Huminerals' Are Extracted, Exploited And Disposed Of*. URL: https://www.forbes.com/sites/jackkelly/2023/02/23/china-workers-have-their-version-of-quiet-quitting-and-acting-your-wage-huminerals-are-extracted-exploited-and-disposed-of/ (visited on 07/24/2023).

Kemalbay, G. and Ö. B. Korkmazoğlu (Jan. 2014). "Categorical Principal Component Logistic Regression: A Case Study for Housing Loan Approval". In: *Procedia - Social and Behavioral Sciences* 109, pp. 730–736.

Kisaog, Z. Ö. (Sept. 2014). "Employee Turnover Prediction using Machine Learning based Methods". Masterarbeit. Ankara, Türkei: Middle East Technical University. URL: https://etd.lib.metu.edu.tr/upload/12617686/index.pdf.

Makaba, T. and E. Dogo (Nov. 2019). "A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms". In: *2019 International Multidisciplinary Information Technology and Engineering Conference*, pp. 1–7.

Najafi-Zangeneh, S. et al. (Jan. 2021). "An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection". In: *Mathematics* 9.11, pp. 1226, 1–14.

Navarra, K. (Apr. 2022). *The Real Costs of Recruitment*. URL: https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/the-real-costs-of-recruitment.aspx (visited on 07/24/2023).

Nguyen, H. M., E. W. Cooper, and K. Kamei (Apr. 2011). "Borderline over-sampling for imbalanced data classification". In: *International Journal of Knowledge Engineering and Soft Data Paradigms* 3.1, pp. 24–29.

Nokeri, T. C. (2021). *Data Science Revealed: With Feature Engineering, Data Visualization, Pipeline Development, and Hyperparameter Tuning*. 1st ed. Berkeley, USA: Apress, p. 22.

Pawar, N., N. Saraf, and T. Pradhan (Jan. 2023). "Role of Analytics in HR-Attrition for effective Decision Making". In: *International Research Journal of Modernization in Engineering Technology and Science*, pp. 185–188.

Raschka, S. (Nov. 10, 2020). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. arXiv: 1811.12808[cs,stat]. URL: http://arxiv.org/abs/1811.12808.

Raza, A. et al. (Jan. 2022). "Predicting Employee Attrition Using Machine Learning Approaches". In: *Applied Sciences* 12.13, pp. 6424, 1–17.

scikit learn (2023). *Glossary of Common Terms and API Elements*. URL: https://scikit-learn.org/stable/glossary.html#term-random_state (visited on 07/24/2023).

Sikaroudi, A., R. Ghousi, and A. Sikaroudi (Oct. 2015). "A data mining approach to employee turnover prediction (case

study: Arak automotive parts manufacturing)". In: *Journal of Industrial and Systems Engineering* 8.4, pp. 106–121.

Soner, S. et al. (Dec. 20, 2022). *Predictive Deep Learning approach of employee attrition for imbalance datasets using SVMSMOTE algorithm with Bias Initializer*. preprint. In Review, pp. 1–19.

Srivastava, P. and P. Eachempati (Nov. 2021). "Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach". In: *Journal of Global Information Management* 29.6, pp. 1–29.

Tharwat, A. (Jan. 2020). "Classification assessment methods". In: *Applied Computing and Informatics* 17.1. Publisher: Emerald Publishing Limited, pp. 168–192.

Vabalas, A. et al. (Nov. 2019). "Machine learning algorithm validation with a limited sample size". In: *PLOS ONE* 14.11, pp. 1–20.

Valle, M. A. and G. A. Ruz (Oct. 2015). "Turnover Prediction in a Call Center: Behavioral Evidence of Loss Aversion using Random Forest and Naïve Bayes Algorithms". In: *Applied Artificial Intelligence* 29, pp. 923–942.

Varma, S. and R. Simon (Feb. 2006). "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7, pp. 91, 1–8.

Verdonck, T. et al. (Aug. 2021). "Special issue on feature engineering editorial". In: *Machine Learning*, pp. 1–12.

Wallace, L. (Mai 2023). *Five Hidden Costs Of Employee Attrition*. URL: https://www.forbes.com/sites/forbeseq/2023/03/21/five-hidden-costs-of-employee-attrition/ (visited on 07/24/2023).