ISSN 2723-6471

Data Processing and Optimization in the Development of Machine Learning Systems: Detailed Requirements Analysis, Model Architecture, and Anti-Data Drift Strategies

Nataliya Boyko[1,*]

[1]*Department of Artificial Intelligence Systems, Lviv Polytechnic National University, 79000, 12 Bandery Str., Lviv, Ukraine*

(Received: May 20, 2024; Revised: June 25, 2024; Accepted: July 10, 2024; Available online: July 31, 2024)

Abstract

The research relevance is determined by the growing need to use machine learning systems in various industries, which requires reliable data processing and optimization. The study aims to develop a machine learning system for data processing and optimization, that predicts employee departure based on internal company data, analyze the subject area and existing approaches, define model architecture and describe the developed system, validate the application's performance on test data, and develop strategies to counteract data drift. To achieve this goal, the applied methods are machine learning algorithms, including, decision tree algorithm, logistic regression, neural networks, and architectural approaches used in machine learning systems with low input data information. This study employs multi-generation model architectures, ensemble methods with LightGBM for robust prediction, and dynamic adaptation strategies to handle feature and data drift. The main results of the study are a machine learning and data pre-processing system for recognizing the risk of employee dismissal, which can serve as a basis for implementing similar services in IT companies. The object of the study is the system of predicting the probability of a particular employee's dismissal within a certain period. It also demonstrates how to cope with all the difficulties of developing a solution based on data of low information content and poor quality. The implemented application, despite the quality of the data and the high imbalance of classes, produces valuable results for the business. The practical significance of this study lies in the possibility of using the developed system to predict and prevent employee losses, which contributes to increasing team stability and improving the efficiency of personnel management, as well as increasing the competitiveness of enterprises.

Keywords: Information Transformation, Resource Efficiency, Time Series, Operationalization, Pre-Processing

1. Introduction

The growing volume of data necessitates accurate processing and optimization of machine learning systems. Detailed requirements analysis and model architecture are crucial for ensuring accuracy and reliability. Anti-data drift strategies are essential for maintaining relevance and resilience against input data changes. This is crucial for successful implementation of machine learning systems in various industries and contributes to the field's development. A machine learning system can help reduce employee turnover, which, particularly in rapidly growing companies, can lead to financial costs and difficulty in finding new staff. Factors such as project, salary, responsibilities, and career direction can influence employee decisions. Preventing these departures can be more cost-effective than constantly hiring new staff. Therefore, developing a system that predicts employee departure risk based on internal company data can solve this problem.

The problem with the study is that the growing amount of available data and the requirements for its processing and modelling create complex challenges. Eliminating data deficits, improving model architecture, and effective strategies to combat data drift are becoming mandatory tasks in the development of machine learning systems. Developing an application for predicting employee resignations in a company also has to address certain issues: uncertainty in data collection and quality, insufficient data information, class imbalance, the need to develop complex architectures, internal restrictions, and privacy.

Previous research in the field of machine learning identified some problems and limitations, but many aspects remain unexplored. As such, Petryna et al. [1] investigated the impact of an iterative method of weighting respondent data on

***Corresponding author: Nataliya Boyko (na_boyko@ukr.net)**
DOI: https://doi.org/10.47738/jads.v5i3.278
This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).
© Authors retain all copyrights

the accuracy of machine learning models for classification tasks [1]. The results show a positive impact on the quality of model training if the data is properly prepared and the variables for weighting are selected. Shumova et al. [2] conducted an experimental study on the use of deep neural networks to generate recommendations for users on the Internet [2]. The results show that the performance of neural systems strongly depends on the type of search model, data quality, and network training methods. Zhuravel et al. [3] investigated the problem of processing and analyzing large amounts of data, which is becoming relevant due to the increased number of information sources [3]. The paper described two paradigms of data processing – streaming and batch – and discussed the technologies required for efficient data processing.

Oliinyk [4] considered the development of a system for analyzing textual data streams [4]. The study covers the system architecture, the mathematical model of data processing, and the use of machine learning to analyze textual data streams. Gamayun et al. [5] considered the improvement of computer-based machine learning systems at the level of the arithmetic-logical framework, in particular, the use of bit-logarithmic data representation for machine learning procedures [5]. The results indicated the potentially high performance of the developed model, especially in the context of neural networks for machine learning. Tomashko [6] explored the principles of applying machine learning to model a logistics complex system [6]. The author discussed the possibility of using artificial intelligence to manage inventory, improve demand forecasting and optimize delivery processes, and presents the business process structure of a logistics management system. Although these studies used various machine learning algorithms and methods, they did not provide specific information on the detailed requirements, model structure, and plan for dealing with changes in input data. The study aims to create a machine-learning system for processing and improving data analysis. The main objectives of the study are to analyze the subject area, existing approaches, and requirements; model architecture and description of the developed system; validation of the application's performance on test data; and strategies to counteract data drift.

2. Materials and Methods

The study utilized machine learning algorithms and architectural approaches to implement a dismissal prediction application. This study utilized machine learning algorithms to analyze and process tabular and temporal data on employee dismissals. These algorithms predict dismissals based on historical salary changes, company tenure, and departmental evaluations. The models are continuously updated to maintain relevance and accuracy, adapting to new data over time. This approach ensures the accuracy and relevance of the forecasts. On the other hand, architectural approaches are crucial for solving problems with low information content in input data and ensuring model stability in changing environments. They help stabilize models and reduce data variable impact on performance. Architectural approaches also enable robust systems to operate under uncertainty. An application analyzing corporate system data was created using these methods, continuously updating to improve predictions. The data-driven system uses machine learning to predict layoffs and builds trends and hypotheses, continuously improving predictions.

The study utilized historical employee data, temporal data, categorical data, feature drift, and data drift analysis to analyze employee satisfaction and turnover. Historical data includes salary changes, temporal data tracks changes over time, and categorical data helps understand departure risk. Feature drift and data drift analysis monitor data distributions for machine learning model accuracy. The study also utilized graphs, cumulative employee distribution functions, visualization to analyze employee distribution, and diagrams of the model's architecture and first-generation architecture. Results included comparisons of tools for detecting data drift, importance of features, and statistical tests of data drift over different time frames. The study used the Python programming language to write code snippets and libraries such as TensorFlow and Keras to develop and train machine learning models. FlowdiaDiagrams and MS Excel were used to create diagrams and graphs, and HoloViews was used to visualize the results. The following parameters were considered: Strategic Business Unit ID (SBUID), MonthToReview, Strategic Business Unit (SBU), and NextCompensationReviewReasonID (figure 1):

ISSN 2723-6471

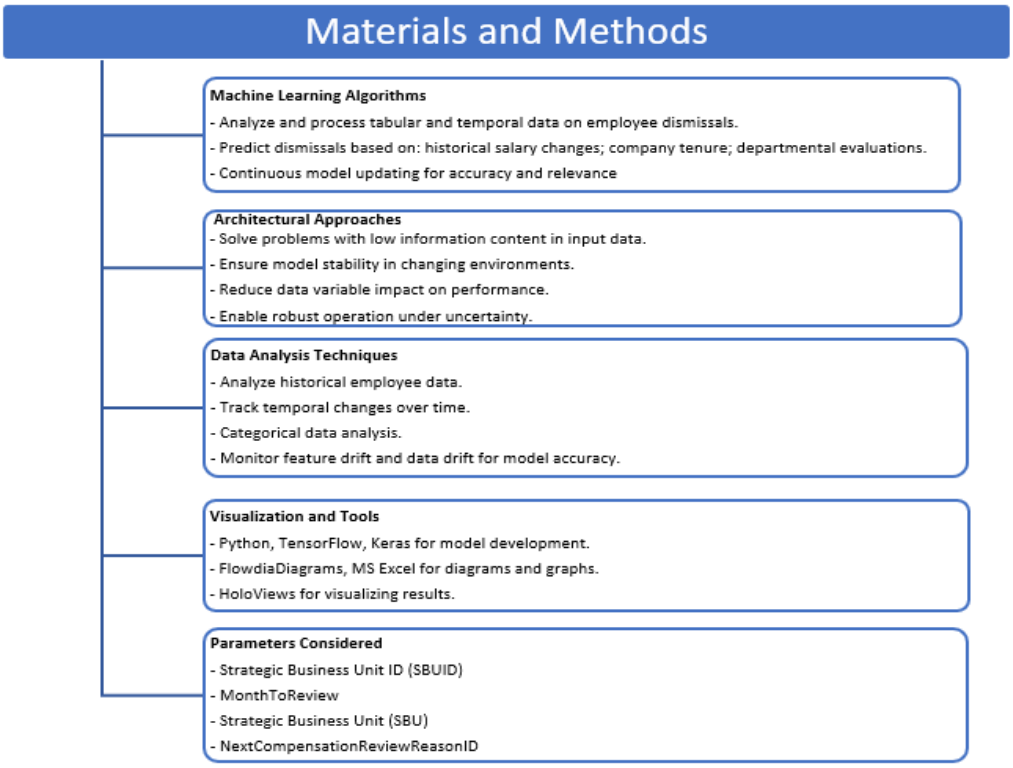

Figure 1. Demonstration of used methods and materials

3. Results

3.1. Analysis of the Subject Area, Existing Approaches, and Requirements

This research combines machine learning and data processing, focusing on large data processing, model optimization, and data drift strategies, requiring a deep understanding of data collection, analysis, and use. Time series forecasting tasks often use time models considering trends and seasonal changes [7]. However, there are two alternative approaches for layoff prediction: statistical-based and dynamic-based. The statistical-based approach uses statistical characteristics to distinguish between dismissal and retention classes, while the dynamic-based approach analyses employee data history to decide whether to terminate them. The assessment's independence from previous data is based on the search for statistical patterns, in particular, using rules within the company that distinguish separate groups of employees based on these patterns. This process can be compared to clustering in classical machine learning, which tries to identify classes based on their clusters rather than individual representatives. The feature drift in the context of retained employees and dismissed employees was analyzed (figure 2).

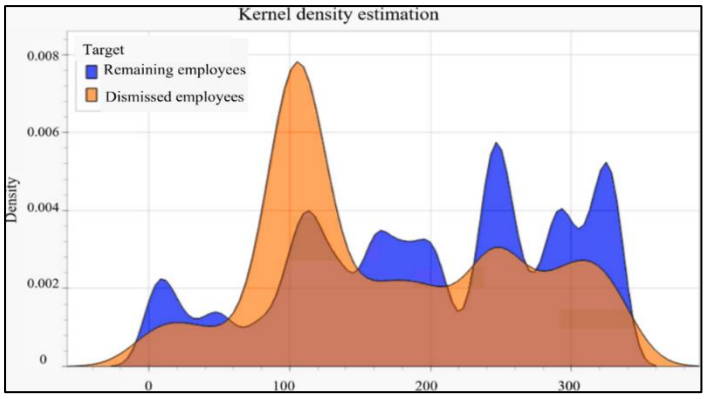

Figure 2. Distribution of dismissed and remaining employees

 ISSN 2723-6471

The figure shows the distribution of dismissed and retained employees based on SBUID within an organization. It uses Kernel Density Estimation (KDE) to illustrate the density of these groups. This figure helps identify patterns in the distribution of dismissed versus retained employees across different departments, enabling targeted interventions and refining predictive models. It is important to note that there is a similar difference in the values of the cumulative function, visually identifying the difference in the distributions when the SBUID peak value is reached. MonthToReview column indicates the number of months left before the employee's salary is reviewed (figure 3).

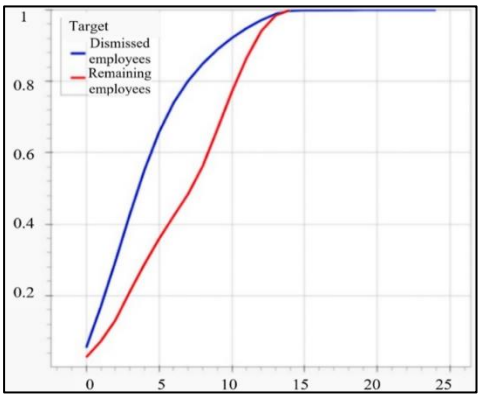

Figure 3. Cumulative function of employee distribution, by MonthToReview column

The figure shows the cumulative distribution function (CDF) of employees based on the "MonthToReview" column, comparing dismissed and retained employees over time. This figure is crucial for understanding temporal patterns of employee turnover, helping identify when employees are more likely to leave the organization and those who stay. From this graph, it is possible to conclude that the dismissed employees have a longer planned time to review their salary. This trend can be seen in the cumulative function plot, where the difference between the distributions is quite significant. Using only statistical rules to divide classes into "resigned" and "working" employees is not an effective method, as this approach does not provide high accuracy and efficiency of the division. On the other hand, the use of decision trees can automatically learn these statistical characteristics of the features without being tied to a history of previous predictions. By analyzing the data as independent records, the true rules and correlations can be more fully explored. And machine learning methods will support statistical methods and allow automated learning of such rules for each feature.

Prediction dependence indicates that when predicting an employee's risk of leaving in month N, the employee's risk is considered, or a vector of features created based on data from month N-1 and all of his or her previous months in the company is used. The prediction formula becomes recursive, where each subsequent prediction depends on the previous one. In the context of this dependency, a "direct dependency" indicates that the input to the risk prediction in month N is based on the feature vector or the prediction itself from month N-1. On the other hand, "indirect dependence" indicates dependence only on the training data related to the previous month (figure 4).

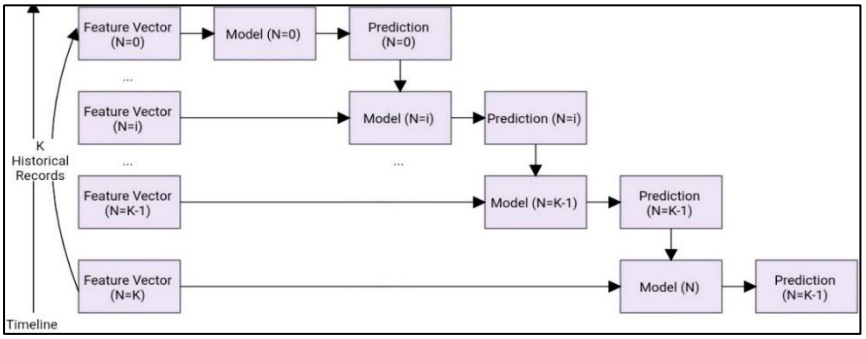

Figure 4. An example of a model architecture with a dependent prediction system

This diagram shows an example of an architecture for predicting employee exit, where each subsequent prediction depends on the previous one. More precisely, "N" is the month number from the beginning of the employee's history in the company; "K" is the number of records about the employee in professional history; "i" is the model iteration

 ISSN 2723-6471

from the first to the last record about the employee in the dataset; “Feature Vector” is a set of features that are fed to the model at the input for making predictions; “Prediction” is the prediction of the employee’s risk of leaving at time “i”. It should be noted that this scheme considers the history of feature vectors and changes in the risk of employee departure, and it is based on text analysis approaches. This study outlines the requirements for a data-driven machine learning system that can predict layoffs, analyze corporate data, build trends, and continuously improve predictions. It emphasizes the importance of large amounts of data, historical data, and internal factors for effective HR processes and informed decision-making. The system also requires regular learning and performance enhancement to ensure high efficiency.

3.3. Model Architecture and Description of the Developed System

To prepare your environment for developing and training machine learning models, as well as for processing and visualizing data, you can use the following code snippets (figure 5):

```
import tensorflow
tensorflow.random.set_seed(1)

from numpy.random import seed
seed(1)

import datetime as dt
from dateutil.relativedelta import relativedelta

import numpy as np
import pandas as pd
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from keras import backend as K
import keras
from keras.models import Model
from keras.layers import Dense, Input, LSTM, Dropout, Bidirectional, TimeDistributed, Flatten, Masking
from keras.optimizers import Adam, SGD
from keras import initializers

import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

Figure 5. Example code for data processing and visualization

This code snippet establishes a robust foundation for developing a sophisticated machine learning system aimed at predicting employee turnover, ensuring reproducibility, effective data handling, and comprehensive model development and evaluation. This part of the code prepares the environment and tools needed for further development, training of machine learning models, and data processing. It is worth analyzing the initial architecture of the developed application model, which can be called the first-generation architecture (figure 6). The main idea of identifying risk areas in this architecture is to divide the model’s predictions into three risk groups: “High Risk”, “Medium Risk”, and “Low Risk”. “High risk” is the smallest group in terms of the number of predictions, includes employees, predicted by 7 or more models, with the highest risk of leaving, a 70% probability. This group emphasizes the importance of the accuracy of predictions. “Medium” risk includes employees with a risk of leaving between 25% and 70% and is predicted by 5 or more models. This group balances between accuracy and completeness of predictions. “Low risk” is the largest group in terms of the number of predictions, includes employees, predicted by fewer than 3 models, with a 25% risk of leaving, focusing on coverage and completeness over accuracy.

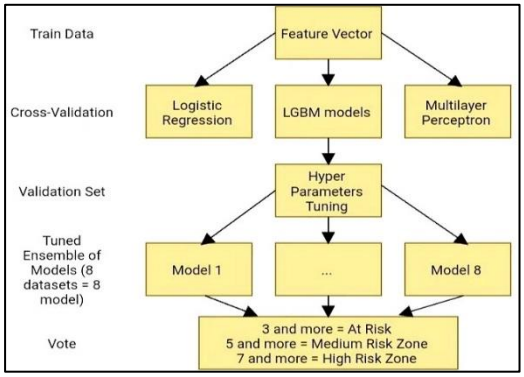

Figure 6. An example of the first-generation architecture in the system

This architecture also has specific steps to determine the risk of dismissal. The first step is "Train Data" i.e., generating and preparing features to create a feature vector. This is followed by "Cross-Validation Training", i.e., training multiple models, such as logistic regression and neural network, to select the best model that considers the balance between accuracy and completeness. The next step is "Hyper-Parameters Tuning", which means tuning the hyper-parameters of the selected model on the validation dataset. Following the hyper-parameters training, the selected model is duplicated (in this case 8 times) and each new model has slightly different prediction parameters to create an ensemble of models. The last step is that the scores of the ensemble of models are aggregated by voting, and the final class is determined based on certain rules, depending on the number of votes for each class.

It should be added that there are certain problems with using neural networks in this system. For example, training recurrent neural networks takes a lot of time due to the individual approach to each prediction. Solutions include data pre-processing, parameter tuning, gradient clipping, regularization, efficient architectures like CNNs, hardware accelerations like GPUs and TPUs, incremental learning, and transfer learning to optimize accuracy and efficiency [8], [9]. Building recurrent models is also a challenge due to the need to select training, validation and test data, and the inadmissibility of losing employees for validation. To ensure the timely quality of predictions, the model must be equally effective for all employees of the company and meet the business requirements. The model may focus on its predictions instead of the input data, which creates an imbalance in the information gain, especially at later stages.

The second-generation application architecture, similar to the first-generation, introduced an ensemble analogue with LightGBM as the base model due to data drift issues, replacing the Cross-Validation Training step. LightGBM is efficient and scalable, handling large datasets and complex models. It can handle categorical features directly, mitigate data drift, and accurately predict minority class instances. It also supports ensemble learning, averaging outputs from multiple models to improve predictive performance. The "Vote" stage was changed: the new rule considered the scores of three models. If two of these models predicted the employee's dismissal, the prediction was considered correct and passed to the next stage. If there were fewer than two votes, the employee was considered to be outside the risk zone. The principle of determining risk zones was also changed: instead of voting models, employees were divided into risk zones based on the critical confidence values of the models. The model provided the probability of dismissal, and the risk zone was determined based on these thresholds. For example, a probability of less than 30% indicated a low risk, up to 50% indicated a medium risk, while a value of more than 75% indicated a high risk.

The third generation of the model can account for employee risk dynamics, analyze statistical characteristics, and use rule-based approaches to further improve the quality of the model. It is important to note that this model is significantly more focused on the first generation than the second, as the first-generation model is more efficient. The third-generation architecture also has "Train Data" (feature generation and feature vector training) and "Classifiers Training" – determining the risk of existing features using LightGBM models for prediction, considering the period during which an employee may leave the company. "Ensemble Votes Aggregation" is a stage that combines the obtained estimates of the risk of employee departure in the previous stage and generates signs to detect changes in the predictions. "Predictions Tuning" – this action includes the selection of hyperparameters to separate risk zones with an ensemble confidence score. "Rebalancing" is the stage responsible for adjusting the results after modelling, using rules to move between risk zones, considering previous and current forecasts. And "Prediction", is when the final values of the employees' assignment to the risk zone are obtained and the corresponding dismissal risk assessment is performed.

3.3. Validation of Application Performance on Test Data

Availability and quality of data are important criteria when developing a machine learning system. Data quality is an attribute that determines the extent to which data is free from anomalies, outliers, input errors or does not meet business standards. This means that the model should be trained on real-world patterns. In this research, various datasets were employed to develop, test, and validate the machine learning models aimed at predicting employee turnover. The datasets included historical records of employees, a validation dataset, a real-world test dataset, and a simulation dataset. The effectiveness of the predictive models was evaluated using a variety of metrics, ensuring a comprehensive assessment of their performance across different aspects.

Various methods should be used to detect anomalies, including statistical, machine learning and rule-based methods. For example, you can check whether salaries are within the acceptable range. Alternatively, you can use more complex

 ISSN 2723-6471

filtering rules, such as filtering salaries by position. The main idea behind data drift detection is to check the consistency of data over time and its quality. This includes preserving the statistical and quantitative characteristics of the data and determining whether the data distributions have not changed with the emergence of new batches of data, which may affect the quality of the model. To detect data drift, author's methods were developed, as the existing packaged solutions did not meet all the needs. For example, different statistical tests are used to detect data drift, depending on the sample size with the same category value (figure 7).

	Great Expectations	Evidently.ai	Whylabs	TFDV
Pandas-based				
standalone tool				
Drift detection				
Big data				
Schema validation				
Data statistics				
Open-source				
Interactive reporting				
Not cloud based				

Figure 7. Comparison of off-the-shelf data drift detection tools

This model is considered to be a good solution for tabular time series problems and meets all the requirements of a particular task. In addition to data drift detection, another important aspect of data validation and processing is feature drift detection. Feature drift is responsible for detecting changes in the statistical characteristics of individual features in datasets. For example, in a predictive modelling scenario for employee retention, feature drift could manifest as changes in the distribution of variables like performance ratings or tenure between training and test datasets. Additionally, feature drift affects predictions by influencing the model's ability to generalize from training to unseen data, leading to reduced accuracy, increased error rates, and diminished reliability in decision-making. Addressing feature drift requires ongoing monitoring and adaptation of models, including regular retraining, feature selection based on relevance, and advanced statistical methods for detecting and mitigating drift. In the case of data drift, statistical differences are usually compared in the time frame. The distribution of individual feature values can change over time or between different parts of the data (figure 8).

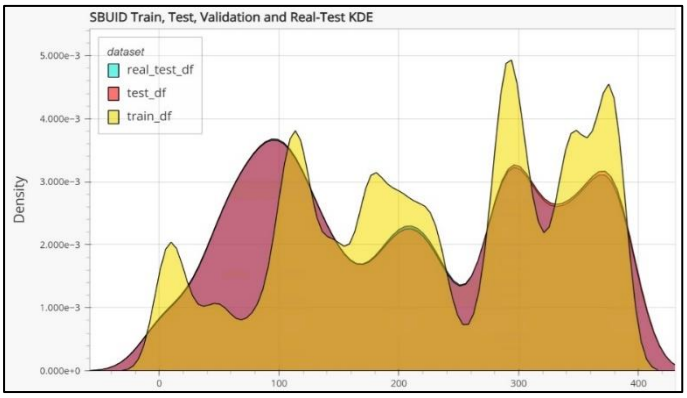

Figure 8. Example of feature drift for the SBUID column

The study reveals a slight shift in features between training and test datasets, but this is considered insignificant due to data temporal distribution. Statistical tests are recommended for accurate assessment of drift. The stability of learned features in training, test, and simulation sets was also investigated to determine if a feature remains relevant and helps distinguish classes. If a feature does not indicate feature drift, retains characteristics on fresh data, and distinguishes classes, it is given high priority in the model. The importance of the attributes should also be visualized using the column "NextCompensationReviewReasonID". Figure 9 is defined for the specified column by target class.

 ISSN 2723-6471

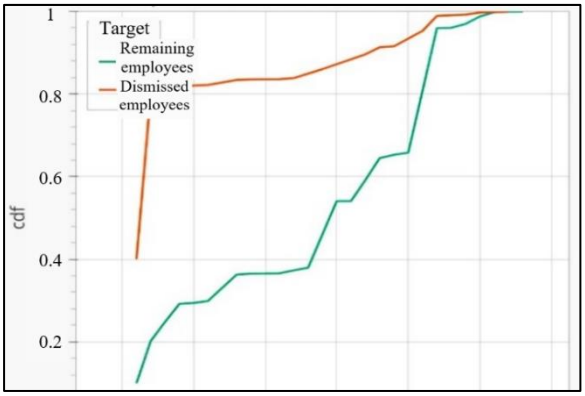

Figure 9. Example of a cumulative function for the NextCompensationReviewReasonID column

The difference in CDF for the two groups of employees is observed. The preservation of statistical features between different data sets is demonstrated. This suggests that the "NextCompensationReviewReasonID" feature is important and should be high on the "feature importance" list for this model. It is also worth discussing this feature separately for the model with the target variable value of 1 (figure 10).

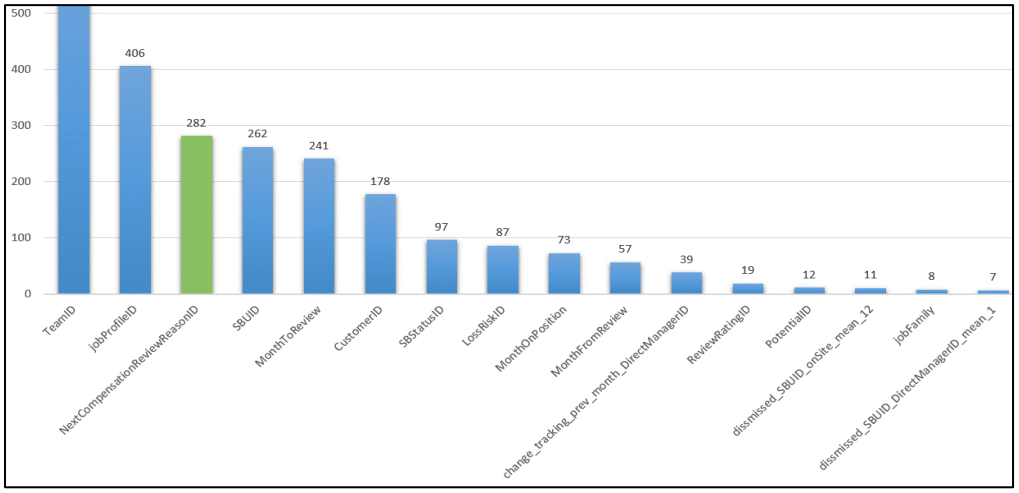

Figure 10. An example of a feature importance list for a LightGBM classifier with target=1

This model is only one of four used at the classification stage. However, the feature "NextCompensationReviewReasonID" remains important regardless of the value of the target variable, which reflects the probability of an employee's resignation. This ensemble of models uses four models with different values of the target variable (1, 2, 3, 4) and aggregates their predictions using a metamodel. Based on the examples above, the determination of feature drift is important for evaluating data and additional features. The importance of "NextCompensationReviewReasonID" is due to its stability in the time aspect and the differences between classes that need to be learned for prediction.

3.4. Strategies to Counter Data Drift

Data drift, a lack of current data, and a lack of helpful features made the programmer challenging to use in forecasting layoffs in the future. The model trained on historical patterns lost relevance, resulting in high performance on training data but poor quality on real data [10], [11]. To address this, it is recommended to avoid using large amounts of historical data and instead focus on newer data, such as training on data from the previous month. The model was designed to capture both long-term and short-term trends, with features including metrics like tenure and recent performance evaluations. Its performance was continuously evaluated using both historical and recent datasets, and hyperparameters were adjusted to balance the influence of both. It is worth considering the list of categorical features that determine data drift (table 1). For statistical analysis, Pearson Chi Squared Test was applied. The Chi-Square test was used in this study to identify shifts in the distribution of categorical variables, such as job roles or departmental assignments, based on observed and anticipated frequencies. By comparing the frequencies of categories in the training

 ISSN 2723-6471

data to those in the new data, the Chi-Square test helped identify significant deviations that might indicate data drift. Differences at $p<0.05$ were considered statistically significant. The elements in which it is detected are marked in red. In other words, the DataDriftDetected column shows whether data drift has been detected. In this case, it is present for more than half of the columns shown.

Table 1. Results of statistical tests of data drift comparing six months of history and the last month

Column	Test	DataDriftDetected	Experiment	P-Value
IsInternalProject	PearsonChiSquared	true	6 Historical Months	0.03
ManagmentLevel	PearsonChiSquared	true	6 Historical Months	0.0
DeliveryGroup	PearsonChiSquared	true	6 Historical Months	0.0
WorkerType	PearsonChiSquared	true	6 Historical Months	0.0
DismissalInitiator	PearsonChiSquared	true	6 Historical Months	0.0
SBStatusID	PearsonChiSquared	true	6 Historical Months	0.0
OnSite	PearsonChiSquared	false	6 Historical Months	1.0
DevCenter	PearsonChiSquared	false	6 Historical Months	0.94
jobFamilyGroup	PearsonChiSquared	false	6 Historical Months	1.0
jobFamily	PearsonChiSquared	false	6 Historical Months	1.0
SBU	PearsonChiSquared	false	6 Historical Months	1.0

An example of a constant drift for the SBU model can also be given (figure 11). The graph shows the change in the number of employees by corporate department compared to the previous month. The absolute values of the differences are hidden due to the sensitivity of the data. The vertical axis shows the difference in the number of employees, and the horizontal axis shows the date against which this difference was calculated for the previous month.

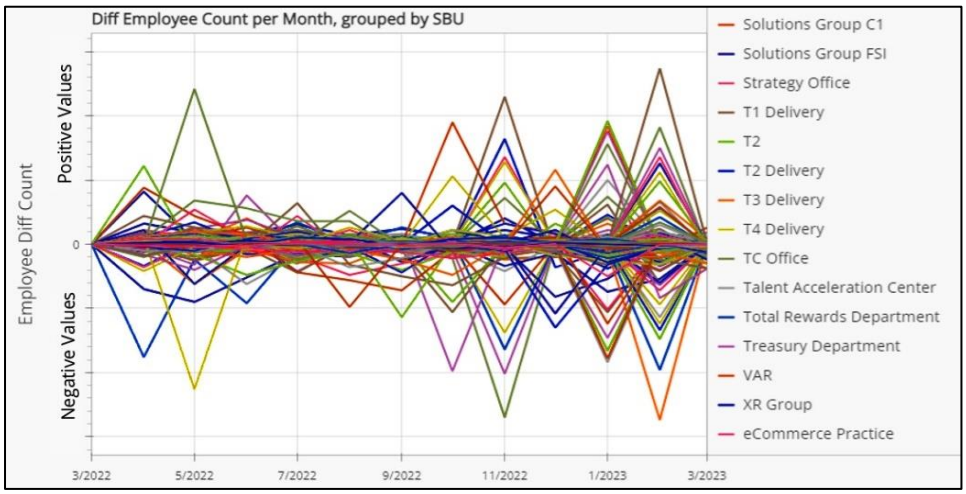

Figure 11. Visualization of the difference in the number of employees by the parameters of the SBU column

It is worth mentioning that neither the statistical nor the dynamic approach solves the problem of data drift. Therefore, it is worth considering certain strategies to counteract it. For example, actively adding new data, as regular data updates allow the model to adapt to changes and improve its relevance, it ensures that the recommendations remain accurate and valuable, even as customer behavior evolves over time. One of the examples is Amazon, which continuously integrates new customer data, such as browsing history, purchase patterns, and product reviews, into its recommendation algorithms. The use of historical data with limited impact, i.e., the ability to train a model on historical data with little impact on current data. Banks and credit agencies use historical data for credit risk assessment, like FICO scores, to reflect current financial behavior. This approach provides accurate assessments, aids lenders in decision-making, and reduces default rates. Use of drift-assisted learning algorithms, i.e., the use of algorithms that are specialized in working with data that changes over time. Google's advertising algorithms utilize drift-assisted learning

ISSN 2723-6471

to adapt to user behavior and market conditions. Reinforcement learning and other algorithms ensure dynamic bidding strategies, targeting criteria, and ad placements, maximizing return on investment. Automatic model adaptation, i.e., a system that automatically analyses data drift and adapts the model to improve its relevance. Netflix uses machine learning models to automatically adjust their recommendation system based on user preferences, ensuring relevance and personalized content suggestions, thereby enhancing user satisfaction and retention. And the use of statistical tests to detect drift, or more specifically, a system that regularly applies statistical tests to detect data drift and respond to it promptly. Regular statistical tests are crucial in healthcare for predictive analytics, detecting drift in patient demographics and health conditions [12]. This allows healthcare providers to update models, maintain high prediction accuracy, allocate resources effectively, and improve patient care outcomes. These strategies can reduce the impact of data drift on the quality of the model and ensure its relevance in a changing environment.

4. Discussion

It is worth mentioning the results of earlier studies. Li et al. [13] devoted their study to the automatic search for optimal data processing sequences for machine learning models. They proposed the DiffPrep method, which provides automation and efficiency in this process. Both studies consider data processing methods for machine learning models. However, the methods themselves are different, as this paper uses machine learning algorithms and architectural approaches, while the others use the DiffPrep method. Oala et al. [14] and Nesterov [15] emphasize that the use of machine learning models in certain industries is limited due to sustainability concerns. One of the main drawbacks is the loss of efficiency due to differences between the data used for training and that used in the working environment. This study shows how these issues can be addressed by combining machine learning robustness testing with physical optics and data-related enhancements to effectively combat drift and provide a dataset. Karl et al. [16] and Lialiuk and Osypenko [17] investigate the various metrics and constraints that create a multiobjective optimization problem. They review the basics of multivariate hyperparameter optimization and demonstrate its importance in machine learning applications by providing an overview of existing optimization strategies. Although the two studies advance machine learning techniques, the differences in focus and methodology are indicative of the various objectives and obstacles found in the fields of predictive modelling and machine learning optimization.

Khaki et al. [18] and Chu et al. [19] proposed the methods to detect unsupervised drifting. The method, proposed by Khaki et al. and Chu et al., uses a two-step approach where the current and initial data distributions are compared and a subset of data that is the root cause of drifting is identified. In the current research, handling drift involves strategies like continuously adding new data, using weighted historical data, and employing techniques such as sliding windows and regular updates, while Khaki et al. [18], and Chu et al. [19], focus on unsupervised methods for detecting drift in textual data, which includes monitoring changes in word usage patterns, semantic shifts, and topic distributions over time. Ahmad and Aziz [20] addressed the problem of detecting anomalies in computer networks and applied data diagnostic methods to classify anomalies. Using preprocessing and feature selection, the work achieved high accuracy on the KDD Cup99 dataset. Both studies consider anomalies and data preprocessing. However, the 2018 paper uses the KDD Cup99 dataset for this purpose, while this paper uses many other datasets [21].

Stüber et al. [22] focused on machine learning and radiomics optimization for predicting overall survival in patients with hepatic metastases from colorectal cancer. The authors developed an extensive experimental radiomics framework for survival analysis, which includes data processing, feature selection, hyperparameter training, and training of various machine learning models. While study by Stüber et al. [22], focuses on radiomics and survival analysis in a medical context with high-dimensional imaging data, the current research addresses predicting employee turnover with temporal, categorical, and numerical data, necessitating different preprocessing techniques, feature engineering methods, and continuous model adaptation strategies. Mishra et al. [23] highlighted the uneven distribution of data. The authors use three different sampling methods Resampling, SpreadSubSampling, and SMOTE to reduce the uneven distribution of data and classify it using the k-nearest neighbours' algorithm. The classification performance is evaluated using various performance metrics. The current research differs from Mishra et al.'s study in handling complex, dynamic datasets and ensuring model relevance. The latter's sampling-based preprocessing approach effectively addresses class imbalance, but the current research requires advanced strategies for detecting and countering data drift, integrating diverse data types, and continuously updating models.

 ISSN 2723-6471

Active learning, a sub-field of machine learning, plays a key role in reducing the requirements for labelled data and facilitates efficient model adaptation across domains by selecting the most informative data for training and model improvement [24]. Salles et al. [25] proposed TSPredIT, provides an integrated framework that combines data preparation and model hyperparameter tuning. The framework is a package for R and allows defining and performing time series forecasting, including data preparation, decomposition, hyperparameter optimization, modelling, forecasting, and accuracy evaluation. However, the current research used different methodologies due to the need to handle diverse data types, address data drift, ensure continuous model updates, and incorporate domain-specific knowledge, which necessitated a tailored approach that goes beyond the proposed integrated tuning framework "TSPredIT".

5. Conclusion

This research aimed to develop a machine learning system for data optimization and processing. It aimed to analyze the requirements of the created application, consider the architecture of the created model, and develop approaches to combat data drift. To achieve the goal, certain architectural approaches and machine learning algorithms were used. The research also included the use of various graphs, diagrams, and code.

The main research results are a system for identifying the potential risk of losing an employee, which can serve as a basis for creating similar services in the field of information technology. The used methodology for predicting employee losses can be an important tool for companies wishing to take preventive measures to retain qualified personnel and reduce the risk of losses. The results also provide an overview of the field of application, existing methods, and criteria. They provide the model structure and characteristics of the developed system. They verify the effectiveness of the application on a test data set and consider approaches to avoid data drift. To improve data accuracy and identify additional relationships, it's crucial to continue analyzing and researching data sources. The developed solution should be applied across various industries to maximize its usefulness. Constant monitoring and adaptation to changes in data and business environment are essential. Involving different departments and professionals in the process will help to understand organizational needs. Machine learning models should be continuously updated to consider new data and business environment changes. Effective data pre-processing techniques can enhance data quality and structure. Employee training is crucial for optimal system use.

The practical application of this study is the ability to use the developed system to predict and avoid employee losses, which contributes to strengthening team stability, improving the efficiency of human resources management, and increasing the competitiveness of enterprises. The main areas for further research in this area include the automation of the data analysis process, the use of deep learning, the application of advanced data processing methods, the integration of data visualization tools, the analysis of the impact of various factors, ensuring data security and confidentiality, the study of new data sources, as well as certain environmental and social aspects.

6. Declarations

6.1. Author Contributions

Conceptualization: N.B.; Methodology: N.B.; Software: N.B.; Validation: N.B.; Formal Analysis: N.B.; Investigation: N.B.; Resources: N.B.; Data Curation: N.B.; Writing Original Draft Preparation: N.B.; Writing Review and Editing: N.B.; Visualization: N.B.; The author has read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The author received no financial support for the research, writing, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The author declares that she has no known competing financial interests or personal relationships that could influence the work presented in this article.

References

[1] V. V. Petryna, A. V. Doroshenko, R. V. Sydorenko and V. M. Teslyuk, "Model and means of data collection and processing using machine learning", *Sci. Bull. UNFU,* vol. 33, no. 3, pp. 102-109, 2023.

[2] L. Shumova, O. Ryazantsev and S. Pokrishka, "Machine learning models for the formation of recommendations", *Bull. Eastern Ukr. Natl. Univ. named after Volodymyr Dahl,* vol. 2, no. 278, pp. 96-105, 2023.

[3] S. Zhuravel, S. Dumych and O. Shpur, "Research of data collection and processing methods in distributed information systems", *Inf. Commun. Technol. Electron. Eng.,* vol. 1, no. 1, pp. 20-38, 2021.

[4] Yu. O. Oliinyk, "Text data stream analysis system", *Examples Math. Model.,* vol. 3, no. 1, pp. 149-158, 2020.

[5] V. P. Gamayun, O. V. Andreev, V. I. Andreev, "Special coding for machine learning systems", *Probl. Inf. Manage.,* vol. 2, no. 70, pp. 24-27, 2022.

[6] A. Tomashko, "Principles of modeling the logistics complex system using machine learning", *Comput.-Integr. Technol.: Educ., Sci., Product.,* vol. 51, no. Jun., pp. 94-100, 2023.

[7] M. Jahin, M. Shovon, J. Shin, I. Ridoy, Y. Tomioka and M. Mridha, "Big data – Supply chain management framework for forecasting: Data preprocessing and machine learning techniques", *Supply Chain Manage.,* vol. 1, no. Jul., pp. 1-26, 2023.

[8] A. Brijith, "Data preprocessing for machine learning", *Int. Center AI Cyber Security Res. Innovations,* vol. 3, no. Oct., pp.1-4, 2023.

[9] A. Amato and V. Di Lecce, "Data preprocessing impact on machine learning algorithm performance", *Open Comput. Sci.,* vol. 13, no. 1, pp. 1-16, 2023.

[10] P. Porwik, K. Wrobel, T. Orczyk and R. Doroz, "FBDD: Feature-Based Drift Detector for batch processing data", *Agricult. Eng,* vol. 2023, no. Sept., pp. 1-24.

[11] G. Rexhaj, "The role of Building Information Modelling in the implementation of sustainable, environmentally friendly, and social infrastructure projects", *Archit. Stud.,* vol. 10, no. 1, pp. 69-78, 2024.

[12] O. Pidpalyi, "Future prospects: AI and machine learning in cloud-based SIP trunking", *Bull. Cherk. State Tech. Univ.,* vol. 29, no. 1, pp. 24-35, 2024.

[13] P. Li, Z. Chen, X. Chu and K. Rong, "DiffPrep: Differentiable data preprocessing pipeline search for learning over tabular data", *Pipeline Eng.,* vol. 1, no. 2, pp. 1-26, 2023.

[14] L. Oala, M. Aversa, G. Nobis, K. Willis, Y. Neuenschwander, M. Buck, C. Matek, J. Extermann, E. Pomarico, W. Samek, R. Murray-Smith, C. Clausen and B. Sanguinetti, "Data models for dataset drift controls in machine learning with optical images", *Transact. Machine Learn. Res.,* vol. 5, no. May, pp. 1-45, 2022.

[15] Nesterov, "Integration of artificial intelligence technologies in data engineering: Challenges and prospects in the modern information environment", *Bull. Cherk. State Tech. Univ.,* vol. 28, no. 4, pp. 82-92, 2023.

[16] F. Karl, T. Pielok, J. Moosbauer, F. Pfisterer, S. Coors, M. Binder, L. Schneider, J. Thomas, J. Richter, M. Lang, E. Garrido-Merchán, J. Branke and B. Bischl, "Multi-objective hyperparameter optimization in machine learning – An overview", *ACM Transactions Evolution. Learn. Optimization,* vol. 4, no. 3, pp. 1-50, 2023.

[17] O. Lialiuk and R. Osypenko, "Features of the implementation of artificial intelligence in construction", *Mod. Technol. Mater. Struct. Constr.,* vol. 35, no. 2, pp. 172-176, 2023.

[18] S. Khaki, A. Aditya, Z. Karnin and L. Ma, "Uncovering drift in textual data: An unsupervised method for detecting and mitigating drift in machine learning models", *Agricult. Eng.,* vol. 2023, no. Sep., pp. 1-8.

 ISSN 2723-6471

[19] R. Chu, P. Jin, H. Qiao and Q. Feng, "Intrusion detection in IoT data streams based on EMNCD with concept drift", *Int. J.Machine Learn. Cybern.,* vol. 2023, no. Oct., pp. 1-25.

[20] T. Ahmad and M. N. Aziz, "Data preprocessing and feature selection for machine learning intrusion detection systems", *ICIC Express Lett.,* vol. 13, no. 2, pp. 93-101, 2018.

[21] B. Bloshchynskyi and Y. Klyatchenko, "Efficiency of computer means for automatic antennas direction in wireless data transmission systems", *Inf. Technol. Comput. Eng.*, vol. 58, no. 3, pp. 33-40, 2023.

[22] A. Stüber, S. Coors, B. Schachtner, T. Weber, D. Rügamer, A. Bender, A. Mittermeier, O. Öcal, M. Seidensticker, J. Ricke, B. Bischl and M. Ingrisch, "A comprehensive machine learning benchmark study for radiomics-based survival analysis of CT imaging data in patients with hepatic metastases of CRC", *Investig. Radiol.,* vol. 58, no. 12, pp. 874-881, 2023.

[23] S. Mishra, P. Mallick, L. Jena and G. S. Chae, "Optimization of skewed data using sampling-based preprocessing approach", *Front. Publ. Health,* vol. 8, no. Jul., pp. 1-7, 2020.

[24] N. Sachaniuk-Kavets'ka, O. Prozor, V. Khomyuk, and R. Shevchuk, "Mathematical description of the differentiation operation in the logical-time environment", *Inf. Technol. Comput. Eng.,* vol. 57, no. 2, pp. 93-98, 2023.

[25] R. Salles, E. Pacitti, E. Bezerra, C. Marques, C. Pacheco, C. Oliveira, F. Porto and E. Ogasawara, "TSPredIT: Integrated tuning of data preprocessing and time series prediction models", *In: Trans. Large-Scale Data- Knowl.-Cent. Syst. LIV, Berlin: Springer,* vol. 14160, no. Sep., pp. 41-55, 2023.