Classification Model for Effective Employee Segmentation

https://doi.org/10.31713/MCIT.2024.013

Bogdan Boiko
Lviv Polytechnic National University
Lviv, Ukraine
b.boiko@ukr.net

Iryna Protcyk
Lviv Polytechnic National University
Lviv, Ukraine
iryna.s.procyk@ukr.net

***Abstract* – In this work, an efficient classification model for staff segmentation is developed. The ensemble is based on machine learning principles, allowing the exploration of the performance of various classification methods and the tuning of hyperparameters to optimize system performance. Additionally, it provides the ability to compare the metric results of trained models, enabling the selection of the best strategy for each problem. The work considers an efficient and automated data processing pipeline, which includes data collection, cleaning, and transformation processes that can be applied in various fields where efficient data processing is required.**

***Keywords* – segmentation; classification methods; ensemble method; machine learning; Ensemble voting method**

I. Introduction

In today's increasingly competitive business world, employee engagement is becoming more challenging due to the difficulty in understanding who your employees are, which employees are beneficial for your business, and what they expect from your company. Additionally, effective advertising campaigns are becoming more expensive, raising the question of how to best allocate funds to capture the attention of staff [1].

Effective employee engagement continuously demands new solutions. Consequently, questions frequently arise regarding the development of new platforms that will help companies better understand newly recruited employees and their categories [2]. This became the primary goal of this work: the creation of a classification model for effectively engaging employees within the company. The subject area of this work explores the potential use of machine learning and data analysis techniques to develop this model, which will enable the classification of personnel [3].

With the increase in retail turnover, the number of business entities rises, leading to heightened competition. Therefore, the main problem this work aims to solve is the need to develop a model that can classify personnel, which will help enterprises better understand their needs in specific categories and provide them with more effective offers [4].

Another widespread problem is the incorrect targeting of advertising campaigns. Many companies often send the same ad to all their staff, regardless of their interest [3]. This can result in a waste of the advertising budget on irrelevant ads, as it will not benefit an uninterested audience.

To optimize the advertising budget, it is crucial to define categories of users and send them targeted advertising. This approach will allow companies not only to save money on advertising but also to improve engagement with less engaged employees by offering a more personalized advertising approach [5].

The task of personnel classification in the context of effective engagement is highly relevant in today's world. While scientific articles and literature touch on this issue, the number of studies that address it in detail using classification methods is limited. However, there are articles and works that partially address this issue [6].

Currently, the amount of data worldwide is growing rapidly, prompting modern companies to actively use big data processing tools to efficiently and reliably manage these information sets. Therefore, developing a model that can be easily integrated with these tools or data processing pipelines is highly relevant today. The created model is easily integrated into existing data processing pipelines or other systems used by companies interested in our product, which makes our work very relevant. Additionally, it adds value by simplifying and automating the worker classification process [2].

II. Methods of Solving

To improve forecasting quality, ensemble voting was used. Ensemble voting is a machine learning method that combines predictions from different models and selects the most popular prediction as the final decision. The main idea of the method is that each model may have its strengths and weaknesses and may contain errors [5]. However, combining their predictions can provide a more accurate and stable forecast. Thus, the ensemble method is a universal solution that accounts for all errors or shortcomings of other models' forecasts to achieve the most effective result.

For a better understanding of the ensemble voting method, let's schematically depict it in Figure 1:

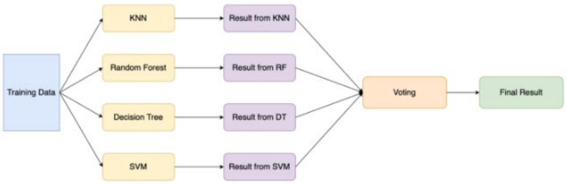

Figure 1. Ensemble Voting Method

As seen in the image, the voting method involves training four separate models on the full set of input data. Each of these models then makes its own predictions. After that, a voting process takes place, and the result that receives the majority vote among the models becomes the final outcome.

III. Results

After creating the ensemble using the following classification methods - K-NN, DT, RF, and SVM- the model was trained. Now, let's take a look at the results of all metrics and graphs for this model (Table 1):

Table 1. Metrics Results for the Ensemble

Metric	Train Result	Test Result
accuracy	0.9876	0.9625
precision	0.9888	0.9631
recall	0.9876	0.9625
F1	0.9852	0.9601

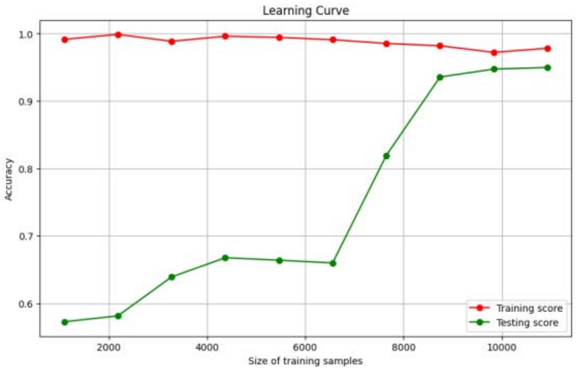

Figure 2. Learning Curve for the Ensemble

By analyzing the results of the metrics and graphs (Figure 2), you can see that this method provides excellent results for our personnel classification task. Indicators such as precision, recall, and F1-score are high, indicating very strong performance from our ensemble.

The learning curve is ascending and free of sudden jumps, which suggests that our training is effective and does not suffer from overfitting or underfitting.

There are a small number of errors in the confusion matrix, but these occur very infrequently, which is acceptable for machine learning tasks. This indicates that our system makes only a minimal number of classification errors, resulting in predictions that are as accurate and efficient as possible.

After a detailed analysis and evaluation of all the results, it was concluded that the ensemble approach, combining K-NN, DT, RF, and SVM, is the most effective solution for the staff segmentation problem.

IV. Conclusion

In today's competitive business environment, employee engagement remains a challenge due to the complexity of understanding employees' needs and expectations. To address this, the development of a classification model aimed at enhancing employee engagement is crucial. The application of machine learning techniques, particularly the ensemble voting method, proved to be highly effective in solving this problem. By combining multiple models, including K-NN, DT, RF, and SVM, the ensemble method produced highly accurate and stable predictions, as reflected in the evaluation metrics.

The results showed high values for accuracy, precision, recall, and F1-score, demonstrating the model's strong performance. The learning curve analysis further supported the model's reliability, as it exhibited consistent training without overfitting or underfitting. The minor classification errors indicated in the confusion matrix were within acceptable limits, further validating the system's efficacy.

In conclusion, the ensemble approach offers a robust and reliable solution for personnel classification, allowing companies to better understand and engage their employees. This approach can also optimize advertising efforts by targeting specific employee categories, resulting in both cost savings and increased engagement. The integration of this model into existing data processing pipelines ensures its relevance and applicability in modern businesses.

References

[1] J. N. Sari, L. E. Nugroho, R. Ferdiana, P. I. Santosa, Review on Customer Segmentation Technique on Ecommerce. Journal of Computational and Theoretical Nanoscience, 2016, 22(10), 3018-3022, https://doi.org/10.1166/asl.2016.7985.

[2] J. Akinsola, Supervised Machine Learning Algorithms: Classification and Comparison. JET Akinsola's Lab, 2017, 48(3), 128 – 138. https://doi.org/10.14445/22312803/IJCTT-V48P126.

[3] A. A. Soofi, A. Awan, Classification Techniques in Machine Learning: Applications and Issues. Journal of Basic & Applied Sciences, 2017, 13, 459-465. https://doi.org/10.6000/1927-5129.2017.13.76

[4] V. Pulabaigari, H.S. Thogarcheti, An Improvement to k- Nearest Neighbor Classifier. IEEE International Conference on Recent Advances in Intelligent Computational Systems (RAICS-2011), 2011. https://doi.org/10.1109/RAICS.2011.6069307

[5] V. Ashok, R. R. Kamath, A. Rk, S. Singh, A. Bhati, Customer Segmentation in E- Commerce. International Journal of Emerging Technologies and Innovative, 2021, 8(7), a908-a913.

[6] R. Entezari-Maleki, A. Rezaei, B. Minaei-Bidgoli, Comparison of Classification Methods Based on the Type of Attributes and Sample Size. Journal of Convergence Information Technology, 2019, 4(3), 94-102. https://doi.org/10.4156/jcit.vol4.issue3.14.