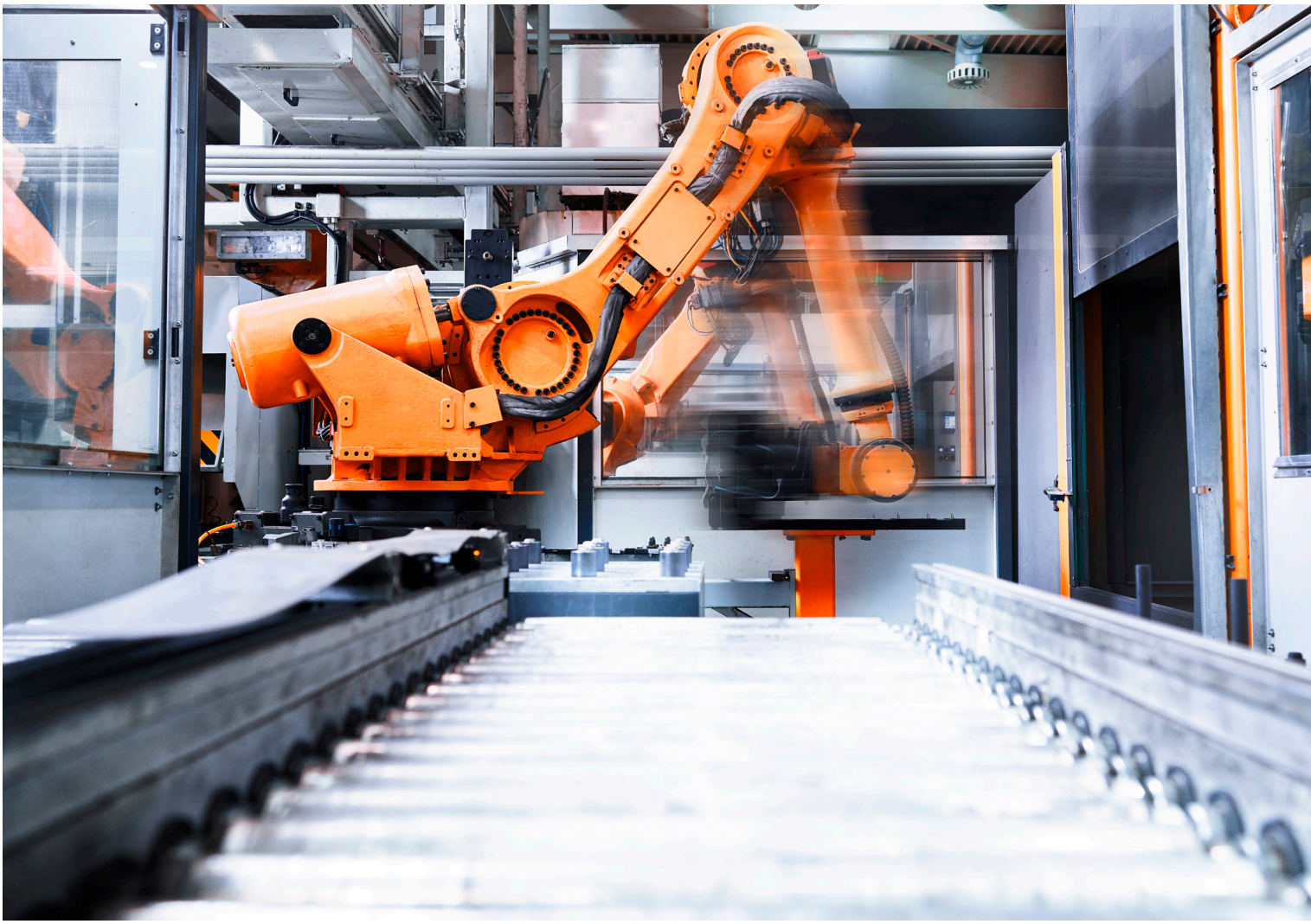
protiviti®
Global Business Consulting



Leveraging Active Learning for Efficient Training of Machine Learning Models with Limited Labeled Data

Face the Future with Confidence®

Abstract

The white paper presents an innovative approach to addressing the challenge of training machine learning (ML) models with limited labeled data using active learning techniques. The primary objective is to reduce the effort and resources required for data labeling, enabling the creation of a data-driven catalog of classified product groups. By utilizing multiple iterations of active learning workflows and incorporating human labeling of a small subset of low-probability records, the proposed methodology aims to mitigate misclassifications which could result in revenue loss due to incorrect discount allocation.

The immediate problem is accurately classifying Stock Keeping Units (SKUs) into appropriate product groups, each associated with a specific discount. The manual misclassifications in the training dataset, stemming from human judgmental errors, further exacerbate the issue. The repercussions of misclassifications include missed opportunities to avail of the correct discounts, potentially leading to revenue losses.

To overcome these challenges, we propose an active learning workflow that optimizes the utilization of labeled data. Active learning allows the ML model to diligently select the most informative and uncertain instances from an unlabeled dataset, requesting human experts to label only those instances which are most likely to improve the model's performance. By iteratively training the model and updating the training set with the newly labeled instances, the model gradually learns to accurately classify product groups, reducing the dependence on large-scale manual labeling efforts.

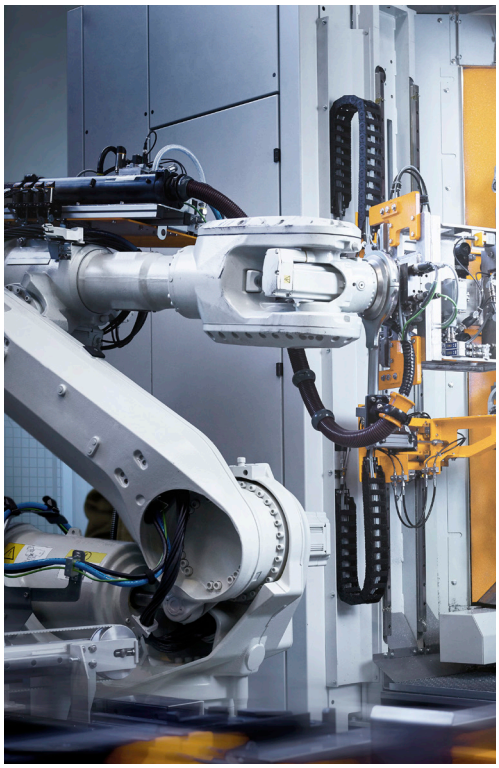
This paper outlines the key components of the proposed active learning workflow and demonstrates its effectiveness through experimental results. We discuss the various strategies employed to select informative instances for labeling and present a comprehensive evaluation of the workflow's performance, including metrics such as classification accuracy, zero one loss and hamming loss, and reduction in manual labeling effort.

The results indicate that the active learning workflow significantly reduces the amount of labeled data required to achieve satisfactory classification performance. By leveraging active learning, the organization can build a data-driven catalog of product groups with associated discounts, minimizing revenue loss due to misclassifications. Further, reducing reliance on manual labeling ensures a more efficient and cost-effective approach to maintaining an accurate and up-to-date classification system.

Introduction

Implementing effective discounts for product groups is critical to determining the success and revenue of a business. However, misclassifications among products, can lead to inaccuracies in discount calculations, resulting in a loss of revenue. Regrettably, the current process of classifying products heavily relies on human judgment and experience alone, which introduces the possibility of inconsistencies in classification outcomes. This lack of consistency poses a significant challenge for future reviews, as there is no reliable and definitive "golden source of truth" for product group classification.

Without establishing a more reliable system, the business faces ongoing risks of revenue loss and potentially enduring consequences. Therefore, it becomes essential to have a classification system that is both consistent and accurate, effectively eliminating the potential for human error and ensuring that discounts are applied correctly. Implementing such a system, will not only prevent revenue loss, but provide a dependable framework for future reviews, enabling the business to make informed decisions and optimize its strategy for product group discounts.



Problem Statement

Discounts based on product groups are essential for businesses, but product misclassification can result in significant revenue losses. However, the current classification process relies heavily on subjective human judgment, leading to inconsistencies and reduced reliability. Moreover, the absence of a definitive "golden source" for future reviews creates downstream issues and further complicates the categorization process. Hence, the goal is to establish a data-driven process to review product categorizations across various SKUs around the globe. The initial phase involves human review, focusing on low-probability classifications to ensure accuracy. As the process stabilizes, these reviewed records are added to the "golden source" for future reference. Ultimately, the aim is to automate the classification process, ensuring consistency and accuracy in categorizations. The business can optimize its revenue and effectively leverage product group discounts by minimizing losses from product misclassification.

Active Learning Methodology

Active learning is a method in machine learning that enhances the learning process by empowering the machine to determine from which data points to learn. It is highly valuable when managing extensive datasets where manually labeling all the data would be time-consuming and costly. By leveraging active learning, the algorithm identifies the most informative data points which require labeling, enabling the machine to learn from a smaller yet highly representative subset. This approach not only saves time and reduces expenses but also enhances the accuracy of the Machine Learning model.

Initialize the Model: Begin by initializing the machine learning model with initial labeled data.

Train the Model: Train the model on the initial labeled data to create a baseline model.

Select Unlabeled Data: From the pool of unlabeled data, select a subset of data points.

Predict Labels: Use the current model to make predictions on the selected unlabeled data points.

Select Most Informative Data Points: Identify the most informative data points based on uncertainty measure (e.g., entropy, confidence, or margin).

Query for Labels: Manually label the selected informative data points through human annotation or expert input.

Augment Labeled Data: Add the newly labeled data points to the initial labeled data.

Retrain the Model: Retrain the model using the augmented labeled data.

Evaluate Model Performance: Assess the model's performance on a validation set or using other evaluation metrics.

Iterate: Repeat the above steps iteratively until the desired performance is achieved.

Experimental Results

The experiment involved a dataset with **7,611** records. The dataset contained features like Supplier, Item Family, Sub-Family, Brand Name, Item Name, and the Product Group. Out of **7,611** records, **761** were labeled. The remaining **6,850** records either had no labels or had labeling issues. Initially, a Multi-class Classification based Random Forest model was trained using 80% of the labeled records and validated on the remaining 20%.

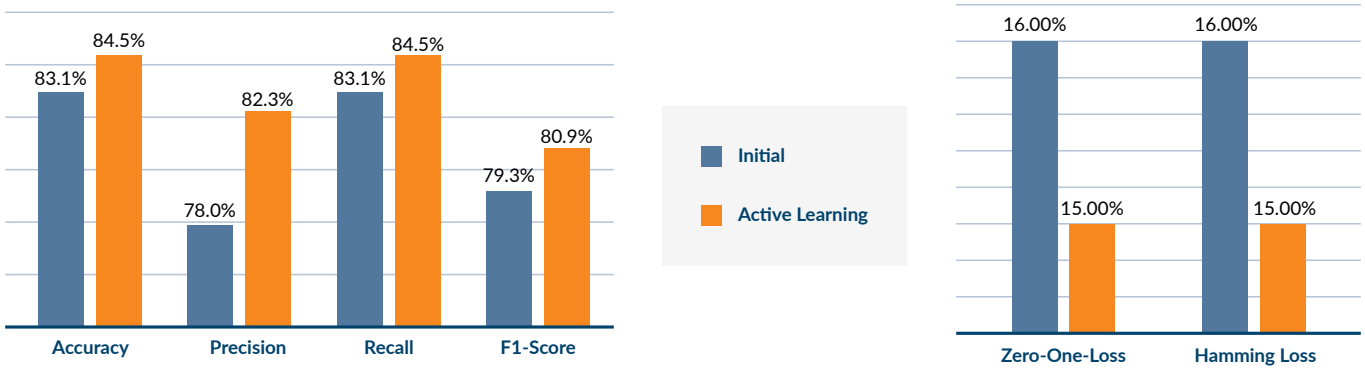
Using this model, the unlabeled records (6,850) were classified. Among these classified records, the model identified the top 50 samples with the lowest confidence, indicating high uncertainty. These samples were then selected for human labeling.

After the human labeling process, the newly labeled records (50 per iteration, 150 in total after 3 iterations) were added back to the original dataset, which originally consisted of 761 records. This loop — involving model training, human labeling, and dataset augmentation — was repeated thrice.

At the end of this Active Learning workflow, the latest model improved significantly in its overall performance, as depicted in Figure 2.

Fig 1: Labeled vs Unlabeled Data Distribution

Fig 2: Comparison of Performance Metrics

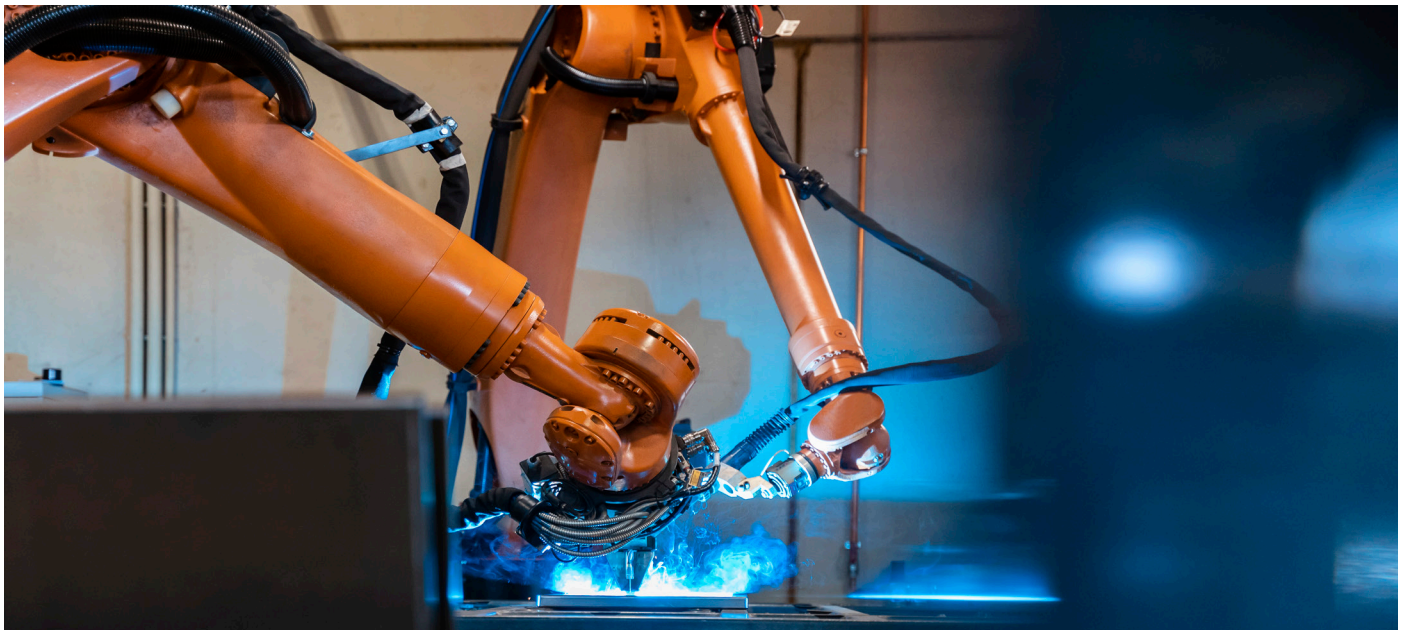


Future Directions

At present, the product group tagging procedure is carried out within the Notebook. However, to enhance user-friendliness, we can introduce a user interface (UI) utilizing an open-source platform like ARGILLA. This implementation would simplify the process of query tagging for individuals (referred to as oracles), allowing interactive engagement. Although we have employed the Random Forest model to address the business problem, there is potential to explore alternative advanced machine learning models which offer improved results and require less time for execution.

Conclusion

By strategically selecting and labeling the most informative instances for training, active learning can significantly improve model performance while minimizing the need for extensive labeled data. It enables efficient data annotation by actively involving human experts in the labeling process, focusing their efforts on the most challenging or uncertain samples. Active learning can also reduce time and cost requirements associated with manual labeling, making it particularly useful in scenarios where labeled data is scarce or expensive to obtain. Overall, active learning empowers machine learning systems to achieve higher accuracy and efficiency by diligently seeking and incorporating targeted information from human annotators.



References

- Settles, Burr. (2012). ***Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning.*** https://doi.org/10.2200/S00429ED1V01Y201207AIM018

- Dasgupta, Sanjoy, & Hsu, Daniel. (2008). ***Hierarchical sampling for active learning. Proceedings of the 25th international conference on Machine Learning*** - ICML '08. https://doi.org/10.1145/1390156.1390216

- Roy, Nicholas, & McCallum, Andrew. (2001). ***Toward optimal active learning through sampling estimation of error reduction. Proceedings of the Eighteenth International Conference on Machine Learning - ICML '01.*** https://doi.org/10.1145/645530.655652

- Tong, Simon, & Koller, Daphne. (2001). ***Support vector machine active learning with applications to text classification.*** Journal of Machine Learning Research, 2, 45-66. http://www.jmlr.org/papers/volume2/tong01a/tong01a.pdf

- Kapoor, Ashish, & Grauman, Kristen. (2012). ***Active learning with Gaussian processes for object categorization. International Journal of Computer Vision, 98(3), 220-240.*** https://doi.org/10.1007/s11263-011-0514-3

- Beygelzimer, Alina, Dasgupta, Sanjoy, & Langford, John. (2009). ***Importance-weighted active learning. Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09.*** https://doi.org/10.1145/1553374.1553497

- Settles, Burr, & Craven, Mark. (2008). ***An analysis of active learning strategies for sequence labeling tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08.*** https://doi.org/10.3115/1613715.1613741

- Guo, Yuxin, Schuurmans, Dale, & Szepesvári, Csaba. (2008). ***Efficient active learning for networked data. Proceedings of the 25th International Conference on Machine Learning - ICML '08.*** https://doi.org/10.1145/1390156.1390221

- Zhang, Tong, et al. (2017). ***Active learning with oracle epiphany. Proceedings of the 34th International Conference on Machine Learning - ICML '17.*** https://proceedings.mlr.press/v70/zhang17a.html

- Sener, Ozan, & Savarese, Silvio. (2018). ***Active learning for convolutional neural networks: A core-set approach. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition - CVPR '18.*** https://doi.org/10.1109/CVPR.2018.00256

- ***Modal A modular active learning framework for Python3:*** https://modal-python.readthedocs.io/en/latest/

- ***Argilla :***https://docs.argilla.io/en/latest/

About Protiviti

Protiviti (www.protiviti.com) is a global consulting firm that delivers deep expertise, objective insights, a tailored approach and unparalleled collaboration to help leaders confidently face the future. Protiviti and its independent and locally owned member firms provide clients with consulting and managed solutions in finance, technology, operations, data, digital, legal, HR, risk and internal audit through a network of more than 90 offices in over 25 countries.

Named to the 2024 Fortune 100 Best Companies to Work For® list for the past 10 years, Protiviti has served more than 80 percent of Fortune 100 and nearly 80 percent of Fortune 500 companies. The firm also works with government agencies and smaller, growing companies, including those looking to go public. Protiviti is a wholly owned subsidiary of Robert Half Inc. (NYSE: RHI). Founded in 1948, Robert Half is a member of the S&P 500 index.

Contact Us:

Amit Lundia
Managing Director
Phone: +91 9836 922 881
Email: amit.lundia@protivitiglobal.in

ACKNOWLEDGEMENT

Data & Digital Solutions Practice
Kallol Kumar
Chiranjib Kashyap Sarma
Bodhisatta Das
Dhairya Lakhani

Designed by
Sweta Roy Choudhury

protiviti®