Generalized Set Segmentation inside ML Training Pipeline for Demand Forecasting

Sokolov Ian
Faculty of Computer Science
National Research University Higher School of Economics
Moscow, Russia
yaosokolov@edu.hse.ru

***Abstract*—The proposed project aims to investigate the effectiveness of generalized set segmentation within a machine learning (ML) training pipeline for demand forecasting. This research will involve a thorough literature review to identify the state-of-the-art techniques and methodologies for data segmentation in demand forecasting. Multiple methods for data segmentation will be implemented and compared to determine the most effective approach. The effectiveness of models trained on segmentated data will also be compared to models trained on the entire dataset. The project will involve the use of ML techniques, statistical analyses, and expert heuristics to evaluate the accuracy and reliability of the models generated. The outcome of this research is expected to improve the accuracy and effectiveness of demand forecasting models, thereby enhancing decision-making in retail industry.**

***Keywords*—Demand forecasting; Machine learning; Data segmentation; Generalized set segmentation; Statistical analysis; Accuracy; Forecasting models; Decision-making; Comparative analysis**

I. Introduction

The field of demand forecasting has become increasingly important in recent years due to the need for accurate predictions of customer demand in various industries. Machine learning (ML) has emerged as a powerful tool for demand forecasting [1, 2, 3], allowing businesses to make informed decisions based on accurate predictions of future demand. However, the effectiveness of ML models depends on the quality of data used in training, including the segmentation of data. In this project, we investigate the use of generalized set segmentation within an ML training pipeline to improve the accuracy of demand forecasting models.

The problem we aim to solve in this research is the challenge of effectively segmenting data for demand forecasting. Traditional methods for data segmentation, such as time-based or cluster-based segmentation, have limitations and may not always be suitable for all datasets. Generalized set segmentation, a novel approach to data segmentation, has shown promising results in improving the accuracy of forecasting models. By investigating the effectiveness of generalized set segmentation within an ML training pipeline, we aim to improve the accuracy and reliability of demand forecasting models, which is critical for decision-making in industries such as retail, logistics, and supply chain management.

To address the challenge of data segmentation in demand forecasting, this research will use a comparative analysis approach to evaluate the effectiveness of generalized set segmentation within an ML training pipeline. We will explore the tradeoff between the limitations posed by small datasets and the presence of logical patterns that can facilitate data segmentation. Multiple methods of data segmentation will be compared, and the accuracy of the resulting models will be assessed to determine the most effective approach. The research will leverage advanced ML techniques and statistical analyses to ensure the accuracy and robustness of the models generated.

In addition, it is important to note that only tree-based ensembles will be used in this research, as they have been shown to perform well in demand forecasting tasks and are robust to noise in the data.

The expected results of this research include the identification of the most effective method of data segmentation for demand forecasting using a comparative analysis approach. By comparing the accuracy of demand forecasting models trained on segmentated data to models trained on the entire dataset, we anticipate that our findings will provide insights into the effectiveness of generalized set segmentation within an ML training pipeline. It is expected that the results of this research will enhance the accuracy and reliability of demand forecasting models, leading to better decision-making process in various industries.

This paper is organized into several sections. The Literature Review section will provide an overview of the relevant literature on demand forecasting, machine learning, and data segmentation. It will review recent advances in demand forecasting techniques, the application of machine learning in demand forecasting, and the use of different methods for data segmentation.

The Methodology section will describe the research approach, including the dataset used, the methods for data segmentation, and the machine learning models trained. The section will also detail the comparative analysis approach that will be used to evaluate the effectiveness of generalized set segmentation within the ML training pipeline.

The Results section will present the findings of the comparative analysis, including the accuracy of the different methods of data segmentation and the performance of the demand forecasting models generated. The section will also provide insights into the effectiveness of generalized set segmentation within an ML training pipeline.

The Conclusion section will summarize the key findings of the research and provide insights into the effectiveness of generalized set segmentation within an ML training pipeline. The section will also discuss the implications of the research findings for businesses and provide recommendations for future research in this area.

II. Literature review

The literature on demand forecasting and machine learning has explored the use of different methods for data segmentation, including the use of a general model trained on the entire dataset versus many specialized models trained on different segments of the dataset. One article [4], concluded that there is no compelling reason to use specialized models since powerful algorithms can natively deal with different behaviors, and using specialized models involves several practical complications. The author compared general and specialized models using 12 real-world datasets with no parameter tuning, and found the general model outperformed the specialized model 89

In this proposed research, we aim to advance the literature on data segmentation by using more advanced clustering techniques to create smarter segmentations. By doing so, we anticipate that our results will demonstrate the effectiveness of generalized set segmentation within an ML training pipeline for demand forecasting, and provide insights into how to improve the accuracy of demand forecasting models. This will represent an advance on what is already known by providing a comparative analysis of the accuracy of demand forecasting models trained on segmentated data versus models trained on the entire dataset.

The next articles [5, 6] examines the impact of segmentation on predictive modeling results. The article suggests that the benefits of segmentation depend on the type of modeling technique used and the specific data being analyzed.

The article highlights that segmentation is beneficial for logistic regression models because they may not pick up on interaction effects between variables. By segmenting the data, these differences can be accounted for, resulting in more accurate predictions. However, for machine learning techniques, the study found that segmentation did not lead to significant improvements in accuracy or lift.

Despite this, the article suggests that there are still benefits to segmenting data for people analytics, including addressing specific areas of interest, accommodating different data, leveraging business understanding, and tailoring interventions. The study concludes that while the impact of segmentation may vary, it is still a useful tool for addressing specific business needs and considerations.

Overall, the article represents an advance in the field by shedding light on the benefits and limitations of segmentation for predictive modeling in people analytics. It provides a nuanced understanding of the role of segmentation in the context of specific modeling techniques and data types, and highlights the importance of considering business needs and interventions when utilizing segmentation.

III. Methodology

The current project aims to investigate the impact of data segmentation on the performance of predictive models using data from the retail, CPG, and supply chain industries. To achieve this, a small dataset consisting of storeId, productId, and demand was used. AgglomerativeClustering algorithm was applied to segment the data, and models were built and trained on both the entire dataset and the segmentated dataset using CatBoost algorithm. The performance of the models was evaluated using the mean squared error (MSE) metric, and it was observed that the models trained on the segmentated dataset provided comparable results to the model trained on the entire dataset.

For future experiments, a real dataset will be analyzed using various plots to understand the data distribution and correlations. Lagged features [7] will be generated, and different clustering algorithms will be tested to identify the best segmentation approach. The experiments will be performed on more powerful machines to tackle the challenges of large data size and computational costs.

The assumption made for this project is that segmenting the dataset may improve the performance of predictive models by capturing the underlying patterns and interactions between the features.

The equipment used for this project includes local machines, and more powerful machines will be searched for future experiments.

To advance the literature further, this proposed research will focus on the development of driver rules for ML pipeline segmentation and will take into account the limitations of the training dataset size. This will contribute to the creation of more effective ML pipelines and improved accuracy in demand forecasting models.

The main difficulties encountered during the project are related to the large size of the data and the computational costs required for the experiments. These difficulties will be addressed in future experiments by using more powerful machines and optimizing the algorithms used.

IV. Results

The initial phase of the study involved gathering a large dataset consisting of 12 million rows with weekly demand, product, and store IDs, as well as independent variables such as promo flags. AgglomerativeClustering algorithm was applied to segment the data, and models were built and trained on both the entire dataset and the segmentated dataset using CatBoost algorithm. Interestingly, the results showed that the models trained on the segmentated dataset provided comparable results to the model trained on the entire dataset. Moreover, a function was developed to calculate lagged features to create more accurate clusters, which could potentially improve the performance of the models. These findings suggest that further investigation into the impact of data segmentation on predictive modeling is warranted.

V. Conclusion

In this work, we explored the potential benefits of segmenting a dataset for predictive modeling. We analyzed a small dataset with only a few features and found that, in some cases, segmentation could result in more accurate predictions. However, this was not always the case, and the benefits of segmentation may depend on the type of algorithm used and the specific characteristics of the data.

Moving forward, we plan to extend our analysis to larger and more complex datasets, including those with time series data. We will also explore more advanced clustering algorithms to determine which techniques work best for different types of data. Additionally, we will create more sophisticated features using lagged data, and investigate the potential benefits of other feature engineering techniques.

Our work has important implications for the field of predictive modeling, as it suggests that segmenting data may be a useful tool in certain situations. By optimizing the use of segmentation and other feature engineering techniques, we may be able to improve the accuracy of predictive models and extract more valuable insights from large datasets.

Overall, our work demonstrates the potential for segmentation to enhance predictive modeling, and we hope that our findings will inspire further research into this important topic.

References

[1] P. Lalou, S. T. Ponis, and O. K. Efthymiou, ”Demand Forecasting of Retail Sales Using Data Analytics and Statistical Programming”

[2] R. Fildes, S. Ma, and S. Kolassa, ”Retail forecasting: Research and practice”

[3] O. B. Yüzbaşıoğlu and H. Küçükaydin, ”Forecasting with Ensemble Methods: An Application Using Fashion Retail Sales Data”

[4] S. Mazzanti, ”What Is Better: One General Model or Many Specialized Models?”.

[5] C. Short, ”Segmentation: Does it Impact Predictive Modelling Results for People Analytics?”

[6] Y. Hadar, ”Should I Train a Model for Each Customer or Use One Model for All of My Customers?”

[7] D. Radečić, ”Time Series From Scratch — Moving Averages (MA) Theory and Implementation”

Word count: 1657