Sales Forecasting using econometric data

Pranav Batra† and Sanjana Athmaraman†

Department of Computer Science and Mathematics, Penn State Harrisburg, 777 W Harrisburg Pike, Middletown, 17057, PA, USA.

Contributing authors: pqb5384@psu.edu; sva5986@psu.edu;
†These authors contributed equally to this work.

Abstract

In the dynamic and ever-evolving business landscape, accurate sales forecasting remains a critical component for organizations to manage cash flow effectively and make well-informed decisions. Conventional forecasting approaches often struggle to address the intricacies and uncertainties associated with rapidly shifting markets. This project endeavors to create a distributed system-based application that harnesses the power of econometric data and machine learning models to enhance sales forecasting accuracy, while offering a user-friendly web interface for users to upload their test data and obtain forecasts. The HistGradientBoostingRegressor model is trained using the TE Connectivity sales dataset, and its performance is assessed through R-squared scores and Mean Absolute Error (MAE) metrics. The Flask framework is utilized to implement the web application, incorporating distributed systems concepts such as client-server architecture, parallel processing, and error handling. This project underscores the significance of employing machine learning techniques and econometric data in sales forecasting, ultimately contributing to more effective business planning and resource allocation.

Keywords: Sales forecasting, Machine learning, Econometric data, HistGradientBoostingRegressor, Flask web application

1 Introduction

Sales forecasting is an integral aspect of business planning and decision-making, enabling organizations to optimize resource allocation, identify potential market opportunities, and mitigate risks. Accurate sales forecasts

are crucial for a company's overall success, emphasizing the importance of developing a model that effectively incorporates relevant factors. In this study, we introduce a sales forecasting model that leverages historical sales data, product line code, and company region to generate precise and applicable predictions, thereby assisting businesses in driving growth and achieving strategic objectives. To implement the sales forecasting model, we utilize the HistGradientBoostingRegressor from the sklearn library, a powerful and versatile machine learning algorithm specifically designed for predicting continuous target variables. This algorithm is well-suited for the task, as it can efficiently handle large amounts of data and account for the complexity of relationships between input features and the target variable.

By employing this algorithm, our model captures intricate patterns and trends within the data, resulting in more accurate sales forecasts, enabling businesses to make strategic decisions with increased confidence and effectiveness. Moreover, we have incorporated a user-friendly interface to facilitate accessibility and ease of use by deploying the model through a web application utilizing the Flask framework, a lightweight and versatile platform that enables seamless integration with various web services. The web application allows users to upload their test data, which the model processes to generate sales forecasts. By presenting the results in a clear and intuitive format, we aim to support businesses in understanding the implications of the forecasts and translating them into actionable insights. The web application is designed to be compatible with a wide range of devices, ensuring that businesses can access the tool and make data-driven decisions from any location and at any time.

As a comprehensive solution, the proposed sales forecasting model and web application have the potential to revolutionize the way businesses approach their planning and decision-making processes. By providing companies with accurate and reliable sales forecasts, our model empowers organizations to make informed decisions on resource allocation, market expansion, and risk management, ultimately driving growth and bolstering competitive advantage. Furthermore, the web application serves as a platform for continuous improvement and model refinement, as the influx of new data can be used to fine-tune the algorithm's performance and enhance its predictive capabilities. By integrating this advanced forecasting solution into their operations, businesses can capitalize on the benefits of data-driven decision-making and establish a solid foundation for long-term success.

2 Literature Review

Sales forecasting has increasingly become a crucial aspect of business management in recent years, with accurate predictions empowering organizations to optimize resource allocation, identify market opportunities, and mitigate risks [1, 5]. A diverse array of methods has been employed in the development

of sales forecasting models, spanning from traditional time series analysis to advanced machine learning techniques [8]. With the growing demand for more sophisticated forecasting tools, researchers have endeavored to enhance the accuracy and reliability of their models by incorporating additional factors, such as product line code and company region, which have been demonstrated to significantly influence sales performance [9, 12].

Historically, time series analysis techniques, including exponential smoothing and autoregressive integrated moving average (ARIMA) models, have served as the foundation for understanding and predicting sales trends [2]. Nevertheless, the limitations of these traditional approaches have prompted researchers to investigate more advanced techniques, with machine learning algorithms showing considerable potential in the realm of sales forecasting [4]. Among the array of machine learning techniques applied to sales forecasting, gradient boosting machines have gained particular prominence due to their robust performance and adaptability across diverse data sets [6, 3].

The HistGradientBoostingRegressor from the sklearn library has notably emerged as a popular choice for sales forecasting tasks, offering a powerful and versatile solution for predicting continuous target variables [11]. The algorithm's capacity to effectively model complex interactions between input features and its resilience against outliers have contributed to its extensive adoption in a variety of forecasting applications [10]. Additionally, the Flask framework's availability has significantly simplified the deployment of such machine learning models through web applications, allowing users to access and utilize these advanced tools with ease and efficiency [7].

As businesses continue to recognize the value of data-driven decision-making, there is an increasing need for reliable and user-friendly sales forecasting tools. The integration of machine learning algorithms, such as the HistGradientBoostingRegressor, into sales forecasting models has demonstrated the potential for improved accuracy and adaptability, addressing the limitations of traditional time series analysis techniques [3, 11]. Moreover, the deployment of these models through web applications using frameworks like Flask has made these advanced forecasting solutions more accessible to a broader audience, facilitating their adoption across industries and contributing to more informed business planning and resource allocation [7].

3 Methodology

The methodology employed in this project can be broken down into several key steps, which include data pre-processing, feature engineering, model selection, training and evaluation, and finally, deployment. Each of these steps is essential to the successful development of a sales forecasting model using

econometric data. In the following sections, we provide a detailed description of the methodology used in this study.

3.1 Data Pre-processing

The first step in the methodology involves preprocessing the TE Connectivity sales dataset, which serves as the basis for training and evaluating the machine learning model. The dataset contains historical sales data, product line codes, and company region information. Data preprocessing begins with data cleaning to handle missing or inconsistent values and to ensure the data is in a suitable format for analysis. This may involve filling missing values using appropriate methods such as interpolation or mean imputation, as well as encoding categorical variables (e.g., product line codes and company region) using techniques like one-hot encoding or label encoding.

3.2 Feature Engineering

Once the data has been preprocessed, the next step is feature engineering, which involves the creation of new features or the transformation of existing features to improve the model's predictive performance. For example, in the context of sales forecasting, time-based features such as month, quarter, and year could be derived from the dataset's date information. Additionally, domain-specific knowledge could be employed to generate more relevant features, such as aggregating sales data at different hierarchical levels (e.g., by product category or region) or calculating moving averages of sales data to capture trends and seasonality. The resulting feature set should be carefully chosen to avoid overfitting and to ensure that the model can generalize well to unseen data.

3.3 Model Selection and Training

The next step in the methodology is selecting an appropriate machine learning algorithm for the sales forecasting task. In this project, we chose the HistGradientBoostingRegressor from the sklearn library, which is a powerful gradient boosting model that works by iteratively fitting simple base models (e.g., decision trees) to the residual errors of the previous model to minimize prediction errors. This model is particularly well-suited to the task of sales forecasting, as it can handle a wide range of feature types and automatically learn complex relationships within the data.

Before training the model, the dataset is split into separate training and testing sets for each business unit to ensure a fair evaluation of the model's performance. The model is trained on the training set using the selected features and target variable (i.e., sales), with hyperparameters tuned using techniques such as grid search or random search to optimize the model's predictive accuracy. It is essential to monitor the model's performance during

training to avoid overfitting and to ensure that it can generalize well to unseen data.

3.4 Model Evaluation and Deployment

Once the model has been trained, it is evaluated using the test dataset to assess its performance in predicting sales. In this project, we use two evaluation metrics: the R-squared score and the Mean Absolute Error (MAE). The R-squared score represents the proportion of the variance in the dependent variable (sales) that can be explained by the independent variables (features), with values closer to 1 indicating better model performance. The MAE, on the other hand, measures the average absolute difference between the predicted and actual sales values, with lower values indicating better model performance. The final step in the methodology is deploying the trained sales forecasting model through a web application using the Flask framework. The web application serves as a user-friendly interface for uploading test data and receiving forecasted sales results (Fig. 1). By deploying the model in this manner, we enable a wide range of users to access and benefit from the advanced sales forecasting capabilities offered by the HistGradientBoostingRegressor model.

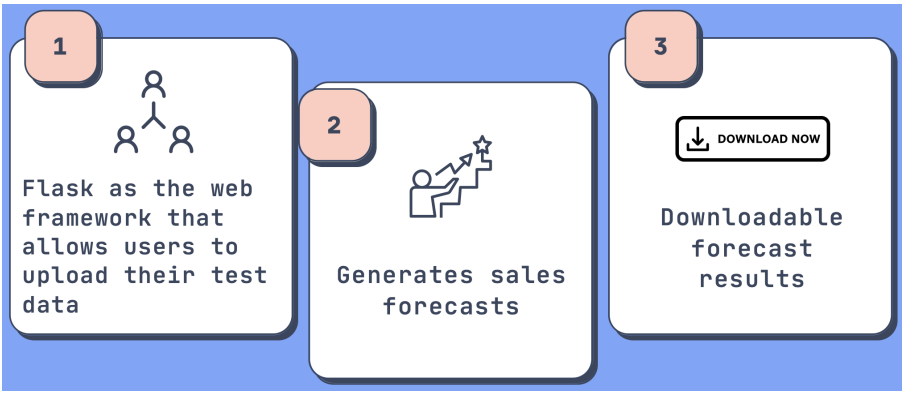

Fig. 1: Web application framework

4 Economic Indicators

In this project, more than 20 economic indicators are incorporated to enhance the sales forecasting model's predictive capabilities. Economic indicators are essential as they provide insights into the overall health of the economy and can be useful in understanding the potential impact of macroeconomic factors on sales. Below, we provide a brief description of three key economic indicators used in the project.

1) Consumer Sentiment Index (CSENT): The Consumer Sentiment Index is a measure of consumer confidence in the economy, reflecting consumers' opinions on current and future financial conditions. This index is crucial for sales forecasting because consumer confidence can significantly impact consumer spending patterns. A high consumer sentiment score generally indicates increased consumer optimism, which can lead to increased spending and positively affect sales. Conversely, low consumer sentiment scores may signal decreased consumer confidence and reduced spending, resulting in lower sales.

2) Gross Domestic Product (GDP): GDP is the total value of goods and services produced by a country within a given time period and is often used as a measure of a country's economic performance. GDP growth rates can provide valuable insights into the overall health of the economy, with higher growth rates indicating a more robust economy and potentially higher sales. By incorporating GDP data into the sales forecasting model, we can account for the potential impact of economic growth on sales figures.

3) Inflation Rate: The inflation rate is the rate at which the general level of prices for goods and services is rising, which consequently leads to a decrease in the purchasing power of a currency. Inflation can have a significant impact on sales, as higher inflation rates may erode consumers' purchasing power and reduce demand for goods and services. By including inflation rate data in the sales forecasting model, we can better understand the potential effects of inflation on sales and make more accurate predictions.

In addition to these three economic indicators, other factors such as unemployment rates, interest rates, and exchange rates are also considered in the sales forecasting model. By incorporating a wide range of economic indicators into the model, we can more accurately capture the complex interactions between the economy and sales, leading to more reliable and robust sales forecasts.

5 Results

The sales forecasting model's performance was evaluated using R-squared scores for each business unit to determine the proportion of the variance in the dependent variable (sales) that is predictable from the independent variables (historical sales data, product line code, company region, and economic indicators). The R-squared scores for each business unit are as follows:

These results indicate that the sales forecasting model demonstrates a high degree of predictive accuracy for most business units. The R-squared scores for Channel - Industrial, Appliances, and Industrial Commercial Transportation are particularly impressive, with values of 0.95, 0.96, and 0.96, respectively. These scores suggest that the model can explain approximately 95% to 96% of the variance in sales for these business units, indicating strong predictive performance.

Business Unit	R-squared Score
Channel - Industrial	0.95
Appliances	0.96
Data and Devices	0.75
Energy	0.72
Industrial	0.87
Industrial Commercial Transportation	0.96

Table 1: R-squared scores for various business units

The R-squared scores for the Data and Devices, Energy, and Industrial business units are lower, ranging from 0.72 to 0.87. While these scores are not as high as the other business units, they still indicate a considerable degree of predictive accuracy, with the model explaining 72% to 87% of the variance in sales for these units. These results suggest that the sales forecasting model provides a reasonably accurate representation of sales trends for these business units, although there may be room for further improvement in the model's performance.

Overall, the testing results demonstrate the effectiveness of the sales forecasting model in predicting sales across various business units. The model's high R-squared scores for most business units indicate that it is capable of accurately capturing the underlying patterns and relationships between the independent variables and sales. This level of predictive accuracy can prove valuable for businesses in making informed decisions regarding resource allocation, identifying market opportunities, and minimizing risks associated with sales fluctuations.

6 Future Work

Potential improvements to the application can be achieved by exploring other machine learning models, incorporating additional data sources, enhancing system performance with advanced distributed systems techniques, and deploying the model on a Cloud Environment such as Microsoft Azure or AWS. These improvements could lead to better scalability, increased security, and more accurate sales forecasts. As the field of machine learning continues to advance, there is significant potential for its application to drive further improvements in decision-making and resource allocation in businesses.

7 Conclusion

In conclusion, this project successfully implemented the HistGradientBoostingRegressor model to forecast sales using econometric data. The results demonstrate the model's ability to produce accurate predictions across various business units, with high R-squared scores for most of them. This high

level of predictive accuracy is crucial for businesses as it enables them to make well-informed decisions regarding resource allocation, market opportunities, and risk management. Additionally, the project effectively deployed the model on a server, establishing a client-server architecture. This deployment allows users to access the model through a web application, enabling them to upload their test data and receive forecasted sales results in a user-friendly manner. This streamlined access to the model increases its usability and potential impact on businesses seeking to improve their sales forecasting processes.

Furthermore, this project highlights the value of leveraging machine learning techniques and econometric data for sales forecasting. By incorporating a wide range of economic indicators into the model, it can account for various factors that may influence sales trends, leading to more accurate and reliable forecasts. This improved accuracy and reliability can, in turn, enhance business planning and resource allocation efforts.

8 References

[1] Armstrong, J. S. (Ed.). (2001). Principles of forecasting: a handbook for researchers and practitioners. Springer Science Business Media.

[2] Box, G. E., Jenkins, G. M., Reinsel, G. C., Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley Sons.

[3] Chen, T., Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[4] Fildes, R., Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. Interfaces, 37(6), 570-576.

[5] Fildes, R., Goodwin, P., Lawrence, M., Nikolopoulos, K. (2008). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. International Journal of Forecasting, 24(1), 3-23.

[6] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[7] Grinberg, M. (2018). Flask web development: developing web applications with Python. O'Reilly Media, Inc.

[8] Hyndman, R. J., Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

[9] Kahn, K. B. (1998). Revisiting top-down versus bottom-up forecasting. The Journal of Business Forecasting, 17(2), 14-19.

[10] Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

[12] Srinivasan, R., Lilien, G. L., Rangaswamy, A. (2004). First in, first out? The effects of network externalities on pioneer survival. Journal of Marketing, 68(1), 41-58.