Mphasis
The Next Applied

Extrapolation Using Regression: Challenges and Solutions

Whitepaper by Manish Shukla, Assistant Manager - Data Science, Mphasis NEXT Labs |
Dr. Archisman Majumdar, Associate Vice President – Applied AI, Mphasis

Contents

Enterprises are often faced with the challenge of predicting the number of complaints and complaint rates for new and yet-to-be-launched products. Often, they have past data on sales, complaints and other attributes available for similar products, but are unable to effectively leverage this information. In this paper, we will discuss the use of Machine Learning models to make predictions for new products, based on reference data of similar products. We will also discuss the challenges in attempting to estimate predictions outside the training data range, and multiple approaches to address them.

1.
Business Use Case and Solution Methodology

Let us take the case of a large global retailer who wants to predict the total number and rate of complaints that their newly-launched product will receive, based on the historical data available for a similar product. The reference product, which has been on the market for some years is chosen as it closely resembles the new product to be launched. A precise prediction of the number of complaints (or the number of complaints per million sales) can help this retailer to efficiently plan human resources and avoid issues of overutilization or underutilization.

To do so, data for sales and complaints should be acquired from several tables, and a data pipeline must be built to handle data cleaning (such as deleting unwanted columns from the data), data aggregation, merging sales and complaints data, and feature engineering. Data pre-processing operations such as scaling and encoding need to be performed to prepare data for model training. For fitting the Machine Learning model, we use AutoML libraries to find the best model with the best parameters. Ordinary Least Square regression (OLS), ridge regression, lasso regression, random forest and XGBoost regression models are fitted, and metrics are compared to find the best model.

All the above models perform well on the test data points within the range of learning data points. Hence, for interpolation, prediction results are very good. However, for data points outside the distribution range of training data, there are no specific methods available to check the accuracy of the regression model. The regression model extends beyond the fitted line for extrapolation or lies out of sample data points. In the absence of knowledge about the real distribution of data in this zone, it is difficult to make predictions with confidence.

2.
Challenges with Extrapolation

The major challenge lies in the region for which predictions have to be made. In the use case cited above, we are trying to predict the number of complaints corresponding to sales numbers that are unseen by the trained model. Hence, it is necessary to extrapolate our results for the new data ranges. Extrapolation is a type of estimation, beyond the original observation range, of the value of

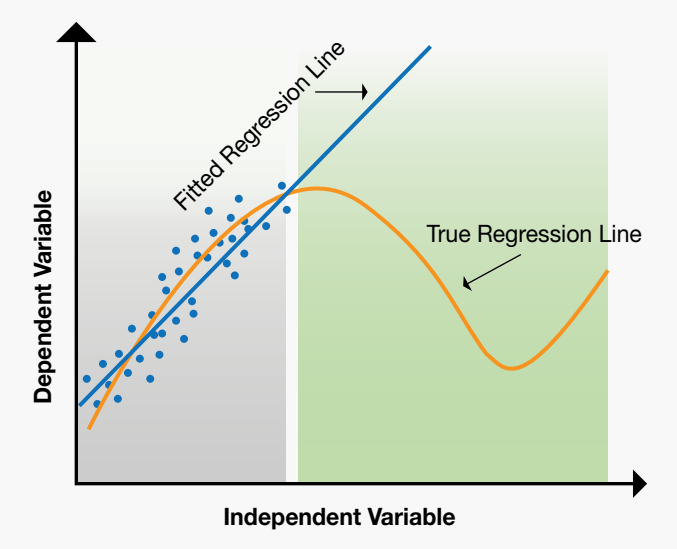

Fig.1: Error in prediction due to extrapolation

a variable, based on its relationship with another variable.

Extrapolation is a valuable method for predicting or forecasting outcomes that are not part of the model's training set. This could include the forecasting size of the population after some years (based on the historical information available on population size and the rate at which population size is changing), anticipating the optimal time to harvest, based on growth rates and weather forecasts and more. Extrapolation becomes less accurate as you go away from the known data or training set, and it becomes even less accurate as the data set's regression expression becomes more complex. If the expected relationship does not hold in the extrapolation zone, the extension of a fitted regression equation beyond the range of the available data might result in seriously skewed results. Even if the relationship is valid, extrapolation can be highly imprecise, even when not biased. This is especially problematic when the extrapolation zone differs significantly from the data region, or if the model is a high-order polynomial or a multiple regression, based on data with high correlation among the independent variables.

Even if a model's fit is excellent, extrapolation outside the data's range should be viewed with caution. Unfortunately and inevitably, extrapolation relies on assumptions about the behavior of the data that cannot be tested beyond their observed support in many circumstances.

Another disadvantage is that extrapolation is not always possible with non-parametric estimating techniques. This is especially obvious while spline smoothing because there are no longer any knots to anchor the fitted spline.

Extrapolation works with linear and other types of regression to some extent, but not with decision trees or random forests. The data is sorted and filtered down into leaf nodes that have no direct relation to other leaf nodes in the tree or forest in the decision tree and random forest, respectively. While the random forest is wonderful for sorting data, the findings cannot be used for extrapolation because it does not know how to classify data outside of the domain.

3.
How Extrapolation Challenges can be Handled?

As discussed in the previous section, extrapolation poses challenges to the accuracy and credibility of the regression model. This section will look at a few approaches that can be used to mitigate them.

Extrapolation necessitates two decisions. First, how valid is the model outside of the data range? Second, how probable is it that a point outside the observed sample range is a member of the population we suppose for the sample?

If these two decisions form contradictory reasoning, we may need to look at other methodologies than extrapolation for better prediction results. On the other hand, if these two decisions endorse the use of extrapolation for making predictions, then the following approaches can be used to improve the accuracy of extrapolation predictions.

The first method is to broaden the prediction region by including upper and lower boundaries in the extrapolation model, rather than providing a single value. While this does not enhance point accuracy, it does enable the forecast to be included in the final score. The actual points that are not on the regression line now fall into the area between the upper and lower limits of the extrapolation. This results in an extrapolation that more accurately includes data points outside of the original set of data. It may not be precise, but it provides enough data to enable planning that considers highs and lows. As we go further away from the known data, this approach gives us a broader range for the predicted outcome, making it worthwhile to keep the data updated as time passes. This, in turn, will help keep the prediction range quite narrow. So, as a rule, avoid extrapolating too far away from the known data range and look at extrapolation predictions of this range with healthy skepticism.

On fitting the model, we have observed that random forest and XGBoost overfit the data. We have thus dropped these models. Among OLS, ridge and lasso, OLS performed the best - with better bias and variance trade-off.

Another way to improve the accuracy of the regression model is to build slightly varied regression relationships and utilize them to generate various extrapolations. These can then be averaged to get the final predictions. It is like creating an ensemble of regression models and taking an average or weighted average prediction. However, this process can be computationally expensive.

4.
Conclusion

Prediction is notoriously imprecise, and accuracy falls as the distance from the learned area grows. If the predictions are to be made beyond the training data range, several checks discussed in this paper must be performed to validate the performance of the model. In situations where extrapolation is required, the model should be updated and retrained to lower the margin of error. Extrapolation is a helpful technique, but it must be used in conjunction with the appropriate model for describing the data, and it has limitations after you leave the training area.

5.
References

1. Gerald J. Hahn (1977) The Hazards of Extrapolation in Regression Analysis, Journal of Quality Technology, 9:4, 159-165, DOI: 10.1080/00224065.1977.11980791
2. https://stats.stackexchange.com/questions/219579/what-is-wrong-with-extrapolation
3. Armstrong J.S. (2001) Combining Forecasts. In: Armstrong J.S. (eds) Principles of Forecasting. International Series in Operations Research & Management Science, vol 30. Springer, Boston, MA. https://doi.org/10.1007/978-0-306-47630-3_19
4. Wing, Coady, and Ricardo A. Bello-Gomez. "Regression discontinuity and beyond: Options for studying external validity in an internally valid design." American Journal of Evaluation 39.1 (2018): 91-108.
5. O'Reilly, Federico J. "On a Criterion for Extrapolation in Normal Regression." The Annals of Statistics 3, no. 1 (1975): 219–22. http://www.jstor.org/stable/2958090.
6. Loh, W.-Y., Chen, C.-W., and Zheng, W. 2007. Extrapolation errors in linear model trees. ACM Trans. Knowl. Discov. Data. 1, 2, Article 6 (August 2007), 17 pages. DOI = 10.1145/1267066.1267067
7. Pomerantsev, Alexey L. "Confidence intervals for nonlinear regression extrapolation." Chemometrics and Intelligent Laboratory Systems 49.1 (1999): 41-48.
8. O'Reilly, Federico J. "On a criterion for extrapolation in normal regression." The Annals of Statistics (1975): 219-222.
9. Malistov, Alexey, and Arseniy Trushin. "Gradient Boosted Trees with Extrapolation." 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE, 2019.
10. Staudenmayer, John, and David Ruppert. "Local polynomial regression and simulation–extrapolation." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66.1 (2004): 17-30.
11. Armstrong, J. Scott. "Forecasting by extrapolation: Conclusions from 25 years of research." Interfaces 14.6 (1984): 52-66

Authors

Manish Shukla

Assistant Manager - Data Science, Mphasis NEXT Labs

Manish Shukla is part of the Mphasis innovation and research group - Mphasis NEXT Labs. He holds an M. Tech. degree from the Indian Institute of Technology, Kanpur in Industrial & Management Engineering. His technical skills include Machine Learning, Deep Learning, Operations Research and Statistical Modeling. He has extensively worked in the field of Text Analytics, Natural Language Processing, Computer Vision, Automated Machine Learning and Information Retrieval from Unstructured Documents. He has a keen interest in advanced AI and ML technologies and has been currently exploring different methods of Applied AI focused on NLP.

Dr. Archisman Majumdar

AVP & Lead - Applied AI, Mphasis NEXT Labs

Dr. Archisman leads a cross-functional team of Data Scientists and consults Fortune 500 companies on AI and ML implementations. He holds a PhD from the Indian Institute of Management Bangalore (IIMB) in the Quantitative Methods and Information Systems area. His areas of expertise are in Machine Learning, Product Management and Information Systems Research.

About Mphasis

Mphasis' purpose is to be the *"Driver in the Driverless Car"* for Global Enterprises by applying next-generation design, architecture and engineering services, to deliver scalable and sustainable software and technology solutions. Customer centricity is foundational to Mphasis, and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization, combined with an integrated sustainability and purpose-led approach across its operations and solutions are key to building strong relationships with marquee clients. Click here to know more. (BSE: 526299; NSE: MPHASIS)

For more information, contact: marketinginfo.m@mphasis.com

USA
Mphasis Corporation
41 Madison Avenue
35th Floor, New York
New York 10010, USA
Tel: +1 (212) 686 6655

UK
Mphasis UK Limited
1 Ropemaker Street, London
EC2Y 9HT, United Kingdom
T : +44 020 7153 1327

INDIA
Mphasis Limited
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village, Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

NR 16/03/22 US LETTER 9ASL7295

www.mphasis.com