Statistical Models vs. Machine Learning: A Comparison in Predictive Analytics

Krishna Mohan Pitchikala

Graduate Student at Department of Computer Science, University of Texas at Dallas, Dallas, TX
Nxp180022@utdallas.edu

Abstract:
Over the past few years, with the rise of the Internet, the amount of data collected from individuals has increased drastically and is anticipated to reach even greater levels in the years to come. To deal with this extensive amount of information, enhanced methods are developed to process and analyze them. With the latest advancements in technologies like machine learning and big data, we can now handle large datasets, clean them, and use them to make predictions based on past experiences. This process is known as predictive analytics, which helps forecast future outcomes using historical data. In prediction analytics and forecasting the results, two commonly used methods are statistical models and machine learning models. Both are aimed at predicting any events or trends that are likely to occur in the future based on the available records, but there are some differences between the approaches. These differences include how many assumptions about the data they make, the degree of easiness of their results in interpretation, their degree of flexibility in the solution of different problems, and their ability to scale. This paper will analyze how statistical models deviate from machine learning models in terms of their merits and demerits. Additionally, it will suggest the best applicable situations for the above methods and analyze relevant literature to show how the two models are used in prediction.

1. Introduction

Data analytics is the process in which raw data is examined, transformed and interpreted with the aim of uncovering patterns, trends and insights that may be useful in decision making. This entails deploying statistical, mathematical and computational techniques to work on data sets to analyze the data in terms of how it may be useful and how it can be changed to provide useful insights for businesses, organizations or researchers. There are four major types of data analytics [1]. So, before we delve into understanding more about predictive analytics, let's take a moment to explore the other types of data analytics to clearly differentiate predictive analytics from the others

1. **Descriptive Analytics:** Descriptive analytics looks at historical data to help understand what happened in the past. It is used to summarize past events using visual plots like line graphs, bar charts, tables, and pie charts, making the information easy to understand. This type of analysis provides a clear picture of trends and patterns over time, allowing businesses to see how customer behaviors or operational processes have changed. By condensing large sets of data into simpler summaries, descriptive analytics offers a high-level view of past performance, helping companies spot cause-and-effect relationships and make sense of their data history.
2. **Diagnostic Analytics:** While descriptive analytics shows what happened, diagnostic analytics aims to identify the reasons behind the occurrence of the event. Its major purpose is to obtain answers regarding why certain events occurred or problems arose. Data mining, data discovery, correlation in data and data transformation are some techniques that facilitate the analysis of data to relay the events that have happened in the past. Diagnostic analytics assists companies in determining which aspects are the source

of the problem assisting more in greening the processes, addressing customer issues, and making more profound decisions to better the processes and performance.

3. **Predictive Analytics:** Predictive analytics on the other hand uses historical data to forecast future trends and events. It is more concerned with describing the likely scenario and the reasons for the event occurrence considering the historical data. Techniques like machine learning, predictive modeling, pattern matching, and forecasting help in the process of identifying the cause-and-effect relationships in data. Predictive analytics will assist businesses in making decisions with high level of awareness cutting down risks, which in turn boosts them in anticipating on issues such as staffing arrangements or stocking of goods appropriate for the demand.
4. **Prescriptive Analytics:** Prescriptive analytics is an extension of predictive analytics. While the predictive analytics forecasts future trends and events, prescriptive analytics prescribes exactly which strategies should be followed for the predicted outcomes to materialize. It assesses various alternatives and their expected consequences and opts for the course of action that enhances doing business processes. This is possible with the employment of machine learning, advanced algorithms, and simulations. With the help of prescriptive analytics, it is possible to better plan and always be one step ahead of the competitors.

Fig.1. Types of data analytics [2]

While all these different data analytics techniques are equally important based on the situation and the information we aim to uncover, in this paper, we will focus more on predictive analytics, with a special emphasis on exploring "How predictive statistical models are compared to machine learning models".

As stated earlier, predictive analysis is a process that uses historical and current data, as well as statistical techniques to forecast events or trends in the future. It includes technologies such as artificial intelligence, data analysis, and predictive analytics to identify relationships in data, which is necessary for forecasting future occurrences. This allows companies and other organizations to make their decisions considerably in advance by predicting what, according to the history, the future behavior of the market or the changes in the individual behavior would be. Predictive analysis dealt with the forth coming event of an event using various statistical as well as machine learning methods. Their outcomes could be the behavior of a consumer that is expected to take place or a change in the market. Predictive analytics enables us to know the probable future events by assessing the past. This approach makes use of many statistical methods, machine learning, predictive modeling and data mining to forecast the future based on the data available in the past and present. In the past, techniques such as linear regression, logistic regression, and time series have dominated predictive analytics. But with the development of new technologies such as decision trees, support vector machines (SVM), and neural networks, new approaches have begun to gain importance. Machine Learning, the processes that enhance the prediction quality by extracting knowledge from past information, has largely

become the focus of predictive modeling. This shift has led to an inquisitive point, which is - whether these machine learning approaches outperform the traditional statistical techniques. To answer this, it is important to understand the fundamental principles behind each approach including their assumptions, interpretability and Flexibility.

2. Statistical Models in Predictive Analytics

Statistical models stand out in forecasting systems as they assist in establishing relationships among data, make extrapolations about trends expected in the future, and assist in decision making. These models enable us to take information from histories and make it into a story worth guessing which the output will be in the future [3].

For instance, in supervised learning, the regression model is able to predict sales in the future dependent on the amount of advertising a business spends and time of the year. Another type, classification models, might be used to predict whether a customer will buy a product or not, based on their past behavior. With the help of clustering models, similar customers can be grouped together thus making it easier for businesses to reach out to them.

These models give the opportunity to the analyst to predict incidents, tune and refine procedures such as increasing efficiency of a production plant, and to even provide assistance in the strategical planning process. Once an analyst is trained sufficiently in the identified techniques, the trust is that he/she will be able to utilize data obtained in making operational and policy decisions of the business or other entities.

3. Machine Learning Models In Predictive Analytics

Machine learning models are very valuable tools for predictive analytics because they are able to find patterns in data and use them for future predictions. With the machine learning models computers can be trained to learn on their own as to what are the expected outcomes based on their previous data. There are various kinds of machine learning models that are suitable for different types of data, and for solving varying problems. More commonly, these models include:

- **Regression analysis:** This is ideal for forecasting or estimating a particular numeric outcome. For instance, one could employ regression in forecasting sales based of advertising spending.
- **Decision trees:** These models assist in performing classification problems, in which there is a need to pick from many different classes. For example, a decision tree model may be designed in determining whether a customer will churn or not based on his or her previous interactions with a company.
- **Support vector machines (SVM):** These models are effective in classification tasks especially if dealing with complex data. For instance, they can be used in classifying emails as spam or not spam.
- **Neural networks:** they are advanced models based on the functioning of the human brain. They are applied for more complicated tasks such as image or voice recognition and even forecasting stock price movements.

4. Key Differences in Statistical Models and Machine Learning Models

4.1. Assumptions and Flexibility:

In almost every statistical method, such as data modeling or hypothesis testing, certain assumptions are made about the data. For example, it might be assumed that the relationship is linear, that error terms are independent, or that the residuals are normally distributed. While these assumptions provide insights into the nature of the data, they also constrain the model's adaptability, especially when the assumptions are found to be false. In contrast, such assumptions are typically weaker in machine learning models, which are more suitable for cases where the relationships between variables are complex and non-linear.

Example from Literature: According to the author of the paper "A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction", machine learning models outperformed traditional statistical methods in terms of flexibility, particularly in datasets with complex interactions and non-linearities [4].

4.2. Interpretability:

As statistical models are based on established statistical theory, they can be easily interpreted. For example, in a linear regression model, the coefficients directly represent the change in the dependent variable for a one-unit change in an independent variable, holding other variables constant. This interpretability is crucial in fields where understanding the relationships between variables is as important as making predictions.

Machine learning models, on the other hand, are known for their accuracy but often sacrifice interpretability. For example, deep learning models are capable of outperforming traditional models in predictive precision, but they do not offer interpretability. These models are particularly difficult to use in high-stakes fields like finance and healthcare, where an explanation of the model behind a decision is a legal requirement. Not many accept this trade-off willingly, and as anticipated, new solutions have emerged, employing measures such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) to address this limitation.

Example from Literature: This issue is also highlighted in the literature, such as in the work titled "Stop Explaining Black-Box Machine Learning Models for High-Stakes Decisions and Use Interpretable Models Instead" [5].

4.3. Scalability and Performance:

The effectiveness of machine learning models typically increases with the size of the training dataset. Neural networks, in most cases, are trained on larger datasets to achieve better performance. The opposite can be said about many statistical models, as their scope is often more limited; some models may not perform well on large datasets due to computational constraints. Therefore, statistical models are often better suited for small datasets, where they offer the additional advantage of providing inferences from the data

Example from Literature: In the book The Elements of Statistical Learning, the authors argue that there is a balance between computational and statistical modeling. They point out that neural network models, as an example of machine learning, work best with large datasets, but that at times, such as when the data does not meet certain conditions, smaller datasets work better with traditional statistical models because the objectives are inference or interpretation and not simply fitting to a complex dataset [6].

4.4. Generalization and Overfitting:

Due to the ineffectiveness of the complexity of statistical models, they are likely to be more over-fitted than under-fitted which relates to their ability to predict the value of an observation not seen by the model. On the other hand, a machine learning or deep learning model can easily do the opposite where it is trained on a small data. To combat overfitting in deep learning, techniques such as dropout, early stopping, and cross-validation are regularly employed, whereas statistical models use metrics like Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) to compare the cost of maximizing the model's fit against the cost of increasing its complexity to gain better results.

Example from Literature: In the book "deep learning" the author reminds us of the dangers of overfitting in neural networks and the usefulness of dropout, early stopping, and cross-validation as regularization techniques.

5. When to Use Statistical Models vs. Machine Learning Models

The appropriateness of statistical and machine learning models is dictated by the problem at hand. Below is a comparative framework elaborating on the model selection issue:

- **Statistical Models**
  - Suitable for small to medium-sized datasets.
  - Used in cases where results need to be explained as logical interactions between various model components.
  - Applicable when the data conforms to accepted distributions and correlations.
  - Suitable for inferential analysis and hypothesis testing.
- **Machine Learning Models**
  - Best suited for complex and large datasets with non-linear relationships.
  - Used when the problem at hand prioritizes prediction over explanation.
  - Suitable for applications where real-time predictions are needed.
  - Effective in cases where the data has high dimensionality or complex feature interactions

6. Conclusion

When it comes to predictive analytics, statistical models and machine learning models have distinct functions, each with their respective advantages and disadvantages. For smaller datasets, statistical models can be clear and useful for inference-based approaches, as long as the assumptions related to the data are valid. The downside is that they may struggle to address large, complex, and non-linear relationships, where machine learning models are often preferred for their superior predictive power, albeit at the cost of interpretability.

A combination of the two approaches seems to represent the future of predictive analytics development. One example is using machine learning models for feature extraction, followed by the application of statistical methods for further interpretation. Explainable neural networks or decision trees may be implemented to achieve interpretability while boosting predictive accuracy.

In the end, the choice between statistical or machine learning approaches is primarily driven by the problem to be solved, the data available, and the intended outcome.

7. References

1. https://www.lotame.com/what-is-data-analytics/
2. https://blog.hubspot.com/marketing/problem-with-predictive-analytics
3. https://graduate.northeastern.edu/resources/statistical-modeling-for-data-analysis/
4. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models by Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B
5. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead by Cynthia Rudin
6. The Elements of Statistical Learning by Hastie, T., Tibshirani, R., & Friedman, J. (2009)
7. Deep Learning by Ian Goodfellow, Yoshua Bengio and Aaron Courville
8. 50 years of Data Science by David Donoho