Proceedings of ICFTBA 2024 Workshop: Finance's Role in the Just Transition
DOI: 10.54254/2754-1169/140/2024.GA18543

The Studies Based on the Application of Predictive Analytics Models

Chenxi Hu[1,a,*], Haiyuying Wen[2]

[1]College of Engineering and Applied Sciences, State University of New York at Stony Brook, Long Island, New York
[2]NingXiang Country Garden School, No.88, Europe South Road, Changsha, Hunan, China
a. chenxi.hu@stonybrook.edu
**corresponding author*

Abstract: The application of predictive analytics models is becoming increasingly prevalent across a range of industries, with relevance extending to numerous sectors. From finance to healthcare to transportation and beyond, the use of predictive analytics is generating significant value. The integration of artificial intelligence techniques, including machine learning (ML), deep learning, and neural networks, has further advanced the capabilities of predictive analytics. In the context of the widespread use of research or work more efficiently, this paper shows that predictive analytics models have significant advantages in weather forecasting, financial analysis, and supply chains. To improve decision-making efficiency, and optimize resource allocation and risk management, this paper systematically reviews and sorts out the papers and research based on predictive analytics models, and deeply excavates the relevance of different models through the analysis of multiple literatures, to help people better choose appropriate models in corresponding scenarios and clarify the role of these models in different industries.

Keywords: forecast model, regression model, time series model, machine learning model.

1. Introduction

Gartner had predicted that “by 2022, over half of major new business systems will incorporate real-time data analytics.” The era of big data is unquestionably upon us. Despite the promise of big data, a 2015 global survey revealed that 43% of companies reported little or no benefit from its use. This is because big data is generated from a variety of diverse channels, and big data is often large and lengthy, complex and variable. People often run into problems when trying to process this data, such as a lack of standards for format, arrangement, quality and security, which makes it difficult for organizations to analyze it effectively. To improve efficiency and accuracy, and not be overwhelmed by a large amount of complex data, entrepreneurs and employees alike seek to break free from these outdated data processing methods. This has driven the creation of models that can manipulate data using smarter means such as ML to create data-driven enterprises.

In this pursuit, the ability to analyze data has become paramount. Predictive analytics models, asset portfolio optimization models, and customer lifetime value models are progressively being integrated into organizations. Among these, predictive analytics stands out for its ability to forecast market trends, assess risks, and predict future outcomes. On the one hand, it can use artificial intelligence

(AI), data mining, ML, modeling, and statistics to make decisive recommendations. On the other hand, by leveraging predictive analytics, companies can better anticipate customer needs, behaviors, and satisfaction, enabling them to devise effective growth strategies swiftly. Predictive analytics is a catalyst for digital transformation in business. In contrast, predictive analytics comprises a range of models, each with distinctive capabilities and applicability to diverse prediction problems. A comprehensive understanding of these models and their operational environments is crucial for enhancing data analysis capabilities and addressing practical challenges.

This paper aims to examine the four branches of predictive analytics models, including regression, time series (TS) models, decision trees, and neural networks. Also, compare the application performance of these models that make up each branch to demonstrate how companies and organizations can more easily benefit from the appropriate models.

2. Forecast Model

“Forecast models are crucial in modern-day business operations, enabling firms to anticipate market changes and adjust strategies accordingly” [1]. Forecast models are essential for predicting short and medium-term trends and influencing future events based on historical data. These models build mathematical frameworks to analyze future trends from incomplete information. They utilize statistical techniques and algorithms to generate forecasts, crucial in fields like finance, weather forecasting, and supply chain management, by identifying patterns in historical data.

Forecast models can be divided into statistical and ML methods. Statistical methods, such as Autoregressive Integrated Moving Averages (ARIMA), rely on mathematical equations to model time-series data. ML methods, like neural networks, use data-driven approaches to learn patterns without explicit programming.

The process of building a forecast model involves several steps. Initially, historical data related to the forecasting problem is collected. This data is then pre-processed to clean and transform it for analysis. Companies need to balance the use of data for productivity with the protection of civilian privacy to avoid legal issues, especially multinational corporations. Next, an appropriate predictive model is chosen based on the data characteristics and prediction goals. The model is then trained using historical data, potentially incorporating ML techniques. Finally, the model's performance is evaluated using indicators like mean absolute error (MAE) and root mean square error (RMSE).

Forecast models provide forward-looking recommendations that enable organizations to make more efficient and orderly decisions. They save time and resources by automating the forecasting process and can handle large datasets, revealing hidden patterns and trends. For example, in finance, predictive models can forecast stock prices and market trends, aiding investment decisions. In supply chain management, they can predict demand, optimize inventory levels, and reduce costs. In a research, Daniel used forecasting models and real-time analytics models to accurately predict near-term traffic volumes, which effectively improved the efficiency of infrastructure such as interstate highways [2].

However, there are challenges associated with forecast models. The complexity of developing and tuning advanced models requires significant expertise. The accuracy of predictions heavily depends on the quality and quantity of historical data; inaccurate or insufficient data can lead to incorrect results. Models may perform well on historical data but fail to generalize to new, unseen data. Additionally, forecast models struggle to account for unexpected events in the long term, and some models, especially those based on ML, can require significant computational resources, impacting efficiency.

The impact of forecast models across various industries is transformative. Before their widespread use, decisions were often reactive and based only on current conditions. Forecast models enable more strategic and long-term planning. For instance, weather forecasting has improved significantly with models like ARIMA for short-term predictions and ML models incorporating multiple variables for long-term forecasts. In research, Mikkel and his team have improved the accuracy of renewable energy forecasts by integrating the intermittency of renewable energy sources such as wind and solar into the energy system using multivariate forecasting models, which shows that these models enhance preparedness for natural disasters and improve planning in agriculture and transportation [3].

In healthcare, forecast models can predict disease outbreaks and patient admissions, helping hospitals allocate resources effectively. In the energy sector, they can predict demand and supply, promoting efficient distribution and reducing costs. Retailers use forecast models to predict product demand across seasons, ensuring optimal inventory levels. In aviation, models predict passenger demand, aiding in optimizing flight schedules and pricing strategies.

In conclusion, forecast models predict future trends based on historical data. Despite challenges like dependency on data quality and computational costs, their impact on various industries is profound. With continuous improvements, especially integrating ML, forecast models will lead to more accurate and reliable decisions.

3. Regression Model

Regression models are fundamental tools in statistical analysis and ML. The core purpose of regression models is to estimate the relationship between the dependent variable (outcome) and one or more independent variables (predictor variables). Regression analysis can be divided into: "logistic regression with the dependent variable as a categorical variable, and Bosson regression with the dependent variable as the count variable" [4]. In univariate linear regression, it can draw a fitting line according to the discriminant function and predict the new data. Of course, multivariate linear regression can also be implemented based on univariate linear regression and machine training, in which case the training data is the number of independent variables. Many people also use first-order curves, second-order curves, and higher-order curves for fitting, which can produce a more approximate effect on the training data, and the expansion of training samples and test samples can achieve small errors and obtain more accurate prediction data. Regression models can also be called a type of ML, after all, every time people input a value to a function, the function will give us the corresponding output value.

When building and applying a regression model, the target website is determined, the research object is first identified, and then the data visualization is carried out, which includes some key points, including having the computer read the data and collect it, cleaning the data, processing the relevant data, dealing with missing values, and transforming the variables if necessary. Then, use visualization tools to visualize your data. Next, it needs to choose the appropriate type of regression model based on the data and research topic. In the process of training the model, it is necessary to use statistical techniques to fit the model to estimate the coefficients, evaluate the performance of the model using metrics such as R-squared and mean square error (MSE), and finally, the model can be used to make predictions on new data.

Regression models have many advantages. It is relatively simple and easy to understand compared to other predictive models, especially linear regression. A large number of predictor variables can be processed at the same time, which greatly improves the efficiency of data processing, and the fixed mathematical construction model also makes many prediction tasks more convenient. Regression analysis can also provide many insights into the relationships between variables, which can be very helpful for making decisions and testing predictions. Despite these advantages, regression models have limitations. That is, predictive models need to assume a linear relationship between the dependent and independent variables, but this is not true in many cases. If the model is too complex, overfitting can occur, which can allow the computer to capture unnecessary values. To better optimize the regression model, more features can be considered as part of the function, and the size of the training and test sets can be increased so that they need to contain more data, which can make the prediction results more accurate.

The regression model makes people's lives more convenient, which can be felt in people's daily lives. Before regression models were widely used, people could only use simple averages to make predictions, which lacked precision. Regression models have made predictions more accurate and reliable, helping many organizations make important strategic plans and enhance operational efficiency. Regression models can be applied to weather forecasting. Univariate linear regression models can even predict temperature based on historical temperature data alone. According to "multivariable regression models significantly improve the accuracy of weather predictions by considering a broader range of influencing factors" [5]. Multivariate linear regression models can provide more accurate predictions by combining additional factors such as humidity, wind speed, and air pressure. In addition, in the financial field, the autoregressive conditional heteroskedasticity (ARCH) model, a branch of the regression model, is used by many companies to analyze and predict market volatility. It is possible to predict the stock index, if the function input is today's stock data information, and the output is the possible stock index information tomorrow. In addition, based on the variance of previous periods, the ARCH model can also predict the variance of the current error. This model is particularly useful in financial markets where volatility is concentrated. "ARCH models provide a robust framework for understanding and forecasting market volatility, thereby aiding in better risk management strategies" [6]. Investors can use the ARCH model to predict stock market volatility and help themselves reduce investment risk more effectively. In conditions that are closer to people's lives, such as shopping platforms, Amazon and Taobao, there are widely automatic recommendation systems on the platform, where the regression model can recommend similar products for the user based on the products that the user is browsing or has browsed in the past, and calculate the likelihood that the user will purchase the product. In general, regression models can analyze variable relationships and make predictions in many scenarios. It is simple, easy to interpret, and computationally efficient. While, at the same time, the premise of its assumptions is objective, and managers need to be careful about data relationships.

4. Time Series Model

TS data consists of observations recorded in a specific time order with consistent intervals between them. When the relationship between variables is indeterminate, TS models are preferred over regression models for prediction. These models rely on temporal dependencies and cannot be predicted from single or independent data points. Popular TS methods include moving average, simple average, and exponential smoothing. The moving average method, effective for eliminating random fluctuations, requires substantial data. Notable TS models include Auto Regressive (AR), Moving Average (MA), ARIMA, and ML models like Long Short-Term Memory (LSTM) networks.

The ARIMA model, powerful in univariate TS forecasting, combines AR, MA, and differencing (I) components. ARIMA is a transformation of ARMA, incorporating AR’s historical value influence, I’s role in converting unstable data to stationary data, and MA’s handling of past random error terms. Exponential Smoothing (ETS) emphasizes recent observations, capturing trends and seasonal changes well. State Space Models, including the Kalman filter and dynamic linear models, use a probabilistic framework for TS data. ML models, such as RNNs and LSTMs, effectively capture time dependencies in sequential data.

TS models offer significant advantages, including adaptability to different types of TS data and improved forecast accuracy through advanced LSTMs. However, they require substantial computational resources, historical data, and expertise. Complex models risk overfitting, and the accuracy and completeness of data are critical. Multivariable data increases the complexity of TS analysis.

TS models have enabled breakthroughs in various fields. In healthcare, they monitor patient vital signs and predict disease outbreaks. In meteorology, spatiotemporal series models extend traditional analysis by incorporating spatial dimensions, essential for weather forecasting. A notable study, "Interpretable weather forecasting for worldwide stations with a unified deep model," demonstrates the use of a unified deep learning model for global weather prediction, leveraging both temporal and spatial data [7]. Research from Tsinghua University further showcases the application of spatiotemporal models in forecasting future weather conditions at meteorological stations, highlighting their ability to interpret complex data [8].

In summary, TS models and their variants offer robust methods for handling complex temporal and spatial patterns, significantly impacting various fields. They enhance understanding of trends, seasonality, and residuals in data, and their influence will continue to grow in the era of big data.

5. ML

The field of ML is a prominent tool in the realm of predictive analytics, which encompasses a diverse array of disciplines. The most common applications of ML for prediction are decision tree algorithms and Artificial Neural Network (ANN).

Decision Tree: This is a technique for estimating a discrete function's value. This is a common categorization technique in which an inductive algorithm is used to provide understandable rules and decision trees after the data has already been analyzed. These choices are then applied to the analysis of the fresh data. There are two separate steps involved in building a decision tree.

Making a decision tree from the training sample set is the first step in the decision tree generating process. A training sample dataset has been processed and analyzed historically, with some degree of synthesis based on real-world data analysis and processing needs.

Pruning decision trees is step two. This step involves testing, adjusting, and fixing the decision tree that was created in the earlier phase. The primary purpose of this step is to validate the basic rules produced during the decision tree generation process using the data in the newly created sample dataset, also referred to as the test dataset. Additionally, it entails cutting off any branches that compromise the pre-balance's precision [9].

Decision trees can play a major role in road accidents. According to the study in, it is shown that when encountering various problems that often occur in actual accident data in traffic accidents, i.e. data integrity and objectivity problems [10]. The decision tree model can better deal with various data noise disturbances in the case of unavailable or irreducible feature information, and to some extent, it can make higher accuracy predictions for classification targets.C5.0 The decision tree method is more robust than other classification models in facing the problems of missing data and high-dimensional input variables, and it can also guarantee a certain degree of classification accuracy.

ANN: A mathematical or computer model that mimics the architecture and operation of a biological neural network—that is, the central nervous system of an animal, especially the brain—is used to estimate or approximate a function in the domains of ML and cognitive science. ANNs with multiple hidden layers demonstrate excellent capabilities for feature learning, whereby the features learned to provide a more fundamental representation of the data, which can be beneficial for visualisation or classification. This paper uses unsupervised learning to accomplish "layer-wise pre-training," a technique that effectively addresses the problem of training deep neural networks [11]. For example, when the TNM staging system is used in conjunction with the ANN system to predict the five-year survival rate of patients with colorectal or breast cancer, respectively, it is shown that the precision of the former is much improved. Furthermore, the versatility of ANNs allows for the incorporation of an array of prognostic factors and the accommodation of continuous variables [12].

6. Conclusion

Based on the review of existing research, this paper presents an analysis of the distinctive characteristics of the four principal predictive analytics models, namely predictive models, regression models, TS models, and ML models, and finds that predictive analytics models have achieved remarkable application results in many fields. The report also examines the comparative advantages of these models over each other and identifies areas where they are better suited for research and development purposes. For example, predictive models are primarily used for weather forecasting, regression models help organizations develop critical strategic plans, time-serial models have gained considerable traction in recent years for macroeconomic control, and ML models are primarily used in transportation, healthcare, and other fields. This advances the understanding of predictive analytics models and sheds light on how the industry can choose models that better align with their needs. However, as data complexity increases and timeliness requirements become more stringent, existing models need to evolve and be optimized. This study not only summarizes the current research progress, but also proposes possible future research directions, such as enhancing the independence of the model, improving the ability of the model to analyze a small amount of data, and optimizing the feasibility of the model. In the current era, people should integrate the research of predictive analytics models with other disciplines and technologies, combined with more advanced artificial intelligence technology and optimized algorithms, to promote the application and development of predictive analytics models in a wider range of fields and make more positive contributions to the development of society and the convenience of people's lives.

Authors Contribution

Chenxi HU and Haiyuying WEN have equal contributions to this article.

References

[1] Yao, D., & Yan, K. (2024). Time series forecasting of stock market indices based on DLWR-LSTM model. Finance Research Letters, 105821.

[2] Wavelet-based denoising for traffic volume time series forecasting with self-organizing neural networks. (2010). Computer-Aided Civil and Infrastructure Engineering, 25(7), 530-545.

[3] Sørensen, M. L., et al. (2022). Recent developments in multivariate wind and solar power forecasting. WIREs Energy and Environment, 12(2).

[4] Mathematical modeling—Predictive models. (2024, August 13). CSDN Blog. https://blog.csdn.net/m0_58585940/article/details/127720448

[5] Daly, K. (2008). Financial volatility: Issues and measuring techniques. Physica A: Statistical Mechanics and its Applications, 387(11), 2377-2393.

[6] Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. Econometrica: Journal of the Econometric Society, 59(2), 347-370.

[7] Wu, H., Zhou, H., Long, M., & Wang, J. (2023). Interpretable weather forecasting for worldwide stations with a unified deep model. Nature Machine Intelligence, 5(6), 602-611.

[8] Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in Neural Information Processing Systems, 34, 22419-22430.

[9] Machine learning and data mining—Classification and predictive models. (2024, August 13). CSDN Blog. https://blog.csdn.net/m0_58585940/article/details/127720448

[10] Candanedo, I. S., Nieves, E. H., González, S. R., Martín, M. T. S., & Briones, A. G. (2018). ML predictive model for Industry 4.0. In Knowledge Management in Organizations, KMO 2018, Communications in Computer and Information Science, 877.

[11] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

[12] Burke, H. B., Goodman, P. H., Rosen, D. B. (1997). Artificial neural networks improve the accuracy of cancer survival prediction. Cancer, 79(4), 857-862.