European Economic Letters
ISSN 2323-5233
Vol 14, Issue 3 (2024)
http://eelet.org.uk

Big Data Analytics in Finance: Predictive Modeling for Investment Strategies

Dr. Vishweswar Sastry V N,
Assistant Professor- Selection Grade, Department of Commerce, Manipal Academy of Higher Education,Manipal.
vishweswar.sastry@manipal.edu

Dr. Guruprasad Desai D.R,
Assistant professor -Selection Grade, Department of Commerce, Manipal Academy of Higher Education, Manipal.
guruprasad.desai@manipal.edu

Dr. Hemanth Kumar
Professor, School of Commerce and Management , Presidency University.
hemanth.kumar@presidencncyuniversity.in

Manjushree M,
Research scholar, School of Commerce and Management, Presidency University,
manju.shree@presidencyuniversity.in

Abstract:

The purpose of this paper is to examine the use of big data analytics in finance based on case study, so our contributions focus predictive modelling for financial investment strategies. It addresses the importance of big data in revolutionizing established financial processes and walks through predictive modeling methods as well with a look at what this means for investment decisions. Finally, we discuss some challenges of big data analytics in finance and provide scope for future research.

Keywords: Big Data, Predictive Modeling, Finance, Investment Strategies, Machine Learning, Data Analytics

1. Introduction

1.1 Background

There has been all the time a notion within the financial industry that may be data used to base their decision upon, but with advent of big say this fact is collected processed and analysed. Historically, financial institutions have utilized historical data and statistical approaches to drive investment decisions such as [1]: But with the volume, variety and velocity of data moving today in a digital way have far exceeded what traditional methods can handle… more advanced ways are required. Pronounce human intelligence is one of the most powerful tools an investor has to help develop accurate hypotheses about risk and return, but it can be augmented by big data analytics – which simply refers to advanced computational analysis using large amounts of disparate information — as well.

Big Data has brought upon a revolution in finance, so why not integrate big data into finance? Incorporating the ability to analyze massive numbers of structured and unstructured data streams in near real-time, financial analysts and investment managers are now able to make better-informed decisions [2]. This is especially true when it comes to predictive modeling: Big data analytics has vastly improved the precision and power of investment strategies. Today, predicting market movements, evaluating risks and optimizing portfolios is a key competitive advantage for any financial institution.

1.2 Problem Statement

Investment strategies can be out-of-date especially when it comes to fast-moving financial markets which snappy trigger trading and arbitraging opportunities in a split second. Historical data and static models form the basis of these strategies, yet they cannot accurately predict the dynamic global markets where information flows continuously and in

unpredictable ways. In addition, the amount of information that can come from different sources such as social networks or news, purchase-sell finance and economic indicators exceed the capabilities of traditional analysis methods [3].

Building predictive models that can use the big data to spot patterns, trends and correlations beyond what a traditional analysis could [do] — in our opinion is where lies the challenge. Such models not only need to process huge data sets in an efficient way, but also they must be able to adapt new information real time and make accurate investment decisions with high returns promptly [4]. With that in mind, it is clear financial institutions are also seeking for new innovative data-driven strategies to drive returns while managing risks amid an increasingly complex market landscape.

1.3 Research Objectives

This paper aims to investigate how predictive modeling is altering investment strategies and why big data analytics has an essential role in this transformation. The aims of this paper are essentially [5]:

- How new and innovative technologies like big data are revolutionizing the finance landscape, particularly with respect to investment strategies.
- Testing the methods and algorithms that work they best at predicting market trends or making investment decisions.
- Surveying how investment strategies using big data analytics perform against traditional approaches.
- Analyzing deception and misinformation in corporate filings
- limitations, ethical concerns and future implications of this area. Social Mill Editor

Future research will explore how the use of big data analytics can improve predictive modeling in finance for more efficient and profitable investment strategies.

2. Literature Review

Big Data came as a great advancement in the financial sector, bringing to life ways to process and analyze large volume of data. Big data in finance is the relationship between huge structured and unstructured sets of information, created from sources such as transactions on markets, consensus economic readings, social networking sites or news plays. Big data has been defined following the 3Vs (volume, velocity and variety) that are out of reach for traditional data processing tools making analytics much more complex [6]. Big Data is applied in various aspects of finance, this includes risk management, fraud detection. customer segmentation and algorithmic trading as well. For example, talk about the way that big data analytics has streamlined risk assessment for financial institutions by allowing them to look at a wider array of mitigating factors and predict potential risks more accurately (e.g. macroeconomic trends or customer behavior changes). Also, feature the aspect of big data in preventing fraud detection – wherever patterns for various...transactions are determined by machine learning algorithms to detect any anomalous activity and thus avoid incidents related with fraudulent actions.

Additionally, big data has revolutionized the trading ecosystem via algo-trading as well ensuring that automated systems execute trades on behalf of traders following pre-defined instructions and real time scrutiny/assessment. As has been emphasized, big data and algorithmic trading are intricately connected with these algorithms being able to analyze market information quickly so as to find high-profitability trades and a little ordering cost. Predictive modeling is predicated upon a foundation of statistical techniques and machine learning algorithms that, effectively crunched through big data analytics, would have the power to anticipate future events based on historical information. Predictive modeling in finance Predictive modeling is one of the most important techniques to develop investment strategies as predictive models helps analysts predict movements and direction of markets, asset prices or risk factors.

There has been a plethora of finance-specific applications for different machine learning models, each with its unique pros and cons. One of the most common statistical methods is regression analysis, which has long been used when predicting asset prices and also market trends. Yet, we know that often regression models may have difficulty in dealing with the complexity and non-linearity of financial data [7], explaining why other more sophisticated techniques such as machine learning are used throughout the industry. Financial predictive modeling has gained popularity mostly in machine learning models such as decision trees, random forests, support vector machines (SVM), and neural networks.

Since these are models that can work with massive datasets and complex inter-variable relationships, they make for good model choices when it comes to predicting markets. Random forests and SVMs have proved to work especially well for classifying financial time series data, whereas neural networks, including deep learning models such as LSTM (Long Short-Term Memory) networks are particularly good at capturing temporal dependencies and trends in sequential data [8].

Finally, deep learning can improve the performance of predicting financial markets and yields good results in predictive modelling. The results show that LSTM networks significantly outpace standard models for predicting stock price behavior due to their ability to learn and remember patterns in long sequences of data [9].

Investment strategies are the practices used by investors and financial institutions to determine where they should allocate their money in order to reach specific investment goals. Historically, traditional investment strategies have been based on tried-and-true data points from the past and fundamental analysis by wall street financial analysts. But the change that has occurred is big data being “the thing” and this introduced a whole new approach to investments, where investment strategies started becoming more Data driven decision making. Quantitative Investing — One of the primary application area for big data in investment strategies is Quantitive investing where complex mathematical models and algorithms are used to exploit market inefficiencies. Quants, as noted above, make use of big data by incorporating vast data streams covering financial statements and macroeconomic indicators along with alternative sources such that social media sentiment. The use of alternative data in investment strategies has been gaining popularity recently. Alternative data includes non-traditional methods like social media posts, web traffic, and satellite images which can offer insights into market trends or company performance that traditional data cannot. Furthermore, the use of alternative from (source) data has become widely accepted to improve model predictions and offer investors better capabilities in shaping their investment strategies more decisively [10].

Lastly, data-driven investment strategies have spawned algorithmic trading systems where relational algorithms execute trades based on real-time analysis of news and market movement. These systems are capable of analysing enormous volumes of data in real time to spot trading opportunities and place orders at speeds that would be unattainable by human traders. As has been noted, algorithmic trading is now a major participant in financial markets globally and represents an ever-increasing piece of total traded volume across all asset types. Big data can offer some compelling benefits for investment strategies, but it's not exactly all wine and roses. Due to complexity of these strategies, it involves most advanced model and need high computational resources are at risk of overfitting i.e. all the models perform well on historical data may not general similarly for future course market conditions as they capture every single minute cases stated in training set. Moreover, the same reliance on automated systems that makes high-frequency trading feasible can bring its own set of new risks – model failures or market manipulation; therefore indicates us to a robust risk-management process.

3. Methodology

The methodology is further provided in the next section, which focuses on how big data analytics can be employed to derive predictive models for investment strategies. This section covers data gathering, predictive modeling methods and process of deployment empowering good foundation for the research.

3.1 Data Collection

These data sources capture some of the complexity inherent to financial markets, which is essential when one strives to deliver well-performing predictive models for investment strategies [11]:

- Past prices for stocks, bonds, commodities and foreign exchange rates from financial databases such as Bloomberg/Thomson Reuters/Yahoo Finance

Macroeconomic data such as GDP growth rates, inflation Rates, unemployment rate and interest rates sourced from government bodies including central banks of various countries or international organization like IMF World Bank. Company financial statement, income statements and other financial reporting from the official source of record for publicly traded companies (E.G. EDGAR – SEC's Electronic Data Gathering and Reporting site). Non-traditional data

sources Like Sentiment scores from social media (Twitter, Reddit etc), news articles, web traffic or satellite imagery via the use of APIs or specialized data wranglers.

Sentiment: Indicates the overall sentiment (positive, negative & Neutral) scores using nlp(Natural language processing) algorithms on posts in social media, news headlines and analyst reports. Pre-processing of the data is very critical because when you clean, modify or convert this real-life dataset to a format that we can work with as columns in pandas Data-frame then only it can be used further for some analysis. The steps are as follows [12]:

Noise removal (unimportant information, duplicates and outliers) Some tools will impute missing data, while others discard the incomplete records altogether. Financial data are frequently at various scales, so they need to be normalized using min-max scaling or Z-score normalization in order for them to fall within the same range. New features from existing data to improve the prediction power of Models. That could be calculating technical indicators (e.g. moving averages, relative strength index) or determining sentiment scores from text data etc.

3.2 Machine Learning Algorithms for Predictive Modeling

This section explains the types of predictive modeling models that applied in this study: Machine learning and deep-learning model [13]. Decision trees are widely used due to their simplicity and interpretability, which allows for identifying important features that drive prices of assets. The study applies common decision tree and ensemble methods such Random Forests to improve prediction accuracy. SVMs are used to classify and predict the movements of asset prices — particularly in cases where variables have a complicated nonlinear relationship. It is bagging based ensemble learning technique that creates multiple decision trees and combine them to get more accurate prediction model or control overfitting. But please note that it works best on large data sets with many features. ANNs are also used to model complex relationships between input variables when the conventional models have difficulty themselves in capturing non-linear patterns. First Feedforward Neural Networks: This study is centered on feed-forward neural networks for a first look.

LSTM addresses the vanishing gradient problem of RNN by utilizing gating designed to allow a cell unit having an option to remember information for long periods, and warehouse it more selectively than before. This concept is especially powerful in the context time-series forecasting, e.g. predicting the stock market one day into the future based on historical data stages. Reasons behind Financial Time Series Data being used in second dimension as 2D Grid are treat it like an image site, CNN is nothing they practically invented for Computer Vision(where there will be lot of local patterns available due to the presence lots of Edge Detection), here we Assume that through this way all primary locations and trends which were not viewable from prior systemic apparatus. Performance evaluation metricsThe performance of predictive models is evaluated using several evaluation metrics:-

• The accuracy rate, which is the proportion of predictions that were correct to all predictions made.

Metrics for evaluating the true positive rate vs. false-positive rate trade-off, particularly crucial in datasets where one class is dominant to others. Harmonic mean of precision and recall -gives a single number that summarizes the overall performance of our model. Calculation of the performance on how well the model can clear out between classes. Scala with higher value represent crisp result. Most often used in regression tasks to quantify the average squared difference between predicted and actual values.

3.3 Implementation

Several predictive models have been implemented using different software tools and programming languages in the study [14]:

- Used for Data preprocessingm model development, and statistical analysis inboth languages. Python libraries are one of the imperative tools for getting started with machine learning and deep learning model implementations like pandas, NumPy, scikit-learn, TensorFlow & Keras.
- To process large amount of data and execute computation on distributed cluster, especially while dealing with massive financial dataset Apache-system is required for preprocessing the information and training the models.

- Use deep learning libraries to create, train and deploy neural networks such as LSTM models for time series analysis.
- For distributed storage and processing of Big Data, especially for unstructured data types like social media feeds or news articles.
- The research follows through a systematic workflow and takes an organized route from implementation of models to data analysis:
- Cleaning, transformation & reshaping data for appropriate analysis.
- Wire store data is pre-processed and predictive models are selected and trained using this processed data.
- Model performance evaluations where appropriate metrics applied and applies cross-validation techniques
- Performing advanced hyperparameters tuning to achieve highest precision and generalization.
- Simulated trading To examine the practical usage of our predictive models on future investment strategies
- Continuous model performance monitoring and adjustments based on market conditions, new data inputs.

Our approach provides a formal and disciplined investigation of big data analytics in predictive modeling for as applied to investment strategies. It guarantees that the research done is grounded in good practices related to data while making use of best tools and techniques necessitated for fair, interpretable work.

4. Case Studies

The next part presents three case studies with applications of big data analytics and predictive modeling in finance, while focusing solely on investment strategies. Case studies on the operational implementation, performance of model and implications for investment decisions are demonstrated in [15].

- In this case study, we look at predicting stock prices and optimizing portfolio allocation in equity markets using predictive modeling. It uses historical price data, trading volumes and other information (financial ratios or sentiment from news), social media posts/stdc.
- The dataset contains daily stock prices of 100 companies from the S&P500 index for the last 10 years, combined with quarterly financial statements and sentiment scores about these companies obtained from social media platforms.
- Models Used • Random Forest classifier to predict stock price direction• LSTM network for future forecasting the next day's closing Osler Shouts
- Data Preprocessing- Noise removal, price normalization and feature engineering (moving averages + sentiment scores) Daily stock movement classifcation (up or down) is performed by the Random Forest model, and LSTM network predicts the closing price as per previous trends.
- The accuracy of the Random Forest model in predicting daily rise/fall was 72% and LSTM had an RMSE to actual_close price of 1.5 %

An algorithmic trading model was then created where we bought stocks that our predictions thought would rise and shorted (i.e., sold) those they believed were part of a downward trend. Overall, it generated an extra 8% a year in excess returns over buy-and-hold. This suggests that using the methodology had potential significant value add to equity investors. The case study demonstrates how big data and predictive modeling help with stock price forecasting and portfolio management. The models were able to give value-added insights for better investment returns by integrating sentiment analysis and financial data. Yet, they also illustrate the risk of model overfitting and argue for regular maintenance to tune models.

The case study is based on using big data analytics in the fixed income market as an example to predict bond prices and model yield curves. To detect defaults a few techniques are used, looking at macroeconomic indicators, credit ratings and interest rate information as some of the examples [16].

This dataset contains the daily bond prices and yields from U.S. Treasury Securities, corporate bonds, and municipal bonds for the past 15 years. It is also provided with the Macroeconomic data (GDP growth Rate, Inflation rates and Central bank policy rate) on priority by Ranking them to higher security space.

The following approach have been devised to predict bond price movement using an Support vector machine(SVM) model and modelling the yield curve based on macroeconomic factors through a multivariate regression analysis.

- Data Preprocessing that consists of cleaning, normalization and feature engineering but mainly concentrates on creating yield spreads and economic indicators as new features. Each of these models: SVM model for price movements and regression model to predict yield curve shifts.
- Within this session, the SVM model resulted in a 68% accuracy predicting bond prices utilizing yield curve data and achieved an R-squared of.85 on historic Information to identify actual relationship-driving behaviour.
- Developed bond trading strategy based on SVM model's predictions long bonds expected to increase, short sell autumnanced decrease. On average, the strategy returned 5% annually with less volatility than an index of the bond market.

In this case study, we applied big data analytics to the fixed income markets focusing on bond price prediction and yield curve analysis using predictive modeling. ResultsIn the results, you can see how combining fundamentals data with machine learning models might help to make sound investment decisions better. But they also argue that market liquidity and credit risk need to be taken into account when developing models.

This case study delves into the utilization of alternative data like social media sentiment, web traffic and satellite imagery to construct predictive models for investment strategies. Instead, attention is paid to how these alternative data sources can reinforce existing financial models. The dataset contains sentiment scores from social media (Twitter), web traffic data for e-commerce companies, and satellite imagery of retail store parking lots spanning the last five years. Convolutional Neural Networks (CNN) for satellite image analysis, Gradient Boosting Machine GBM to predict stock return movements from social media sentiment and web traffic data

This includes preprocessing of satellite imagery to make it amenable for analysis by neural network models; normalization traffic data from the web and quantify social media mood as features. The models made use of both traditional price information as well as other alternative data sources to predict the stock prices. The CNN model has found interesting and accurate ways in which quarterly sales figures are correlated with the usage of parking lots at retail stores, while GBM was able to obtain an accuracy level slightly exceeding 75% for predicting stock price movements based on sentiment data covered by web traffic. We created a investment strategy by combining the forecasts based on GBM model and traditional financial analysis. This strategy beat the S&P 500 by 10% per year, with a vast majority of profits coming from alternative data-driven insights.

5. Results and Discussion

Results—Testimonial of Case Studies How Big Data Analytics and Predictive Modeling Affect the Development of Investment Strategies In the discussion, we explore implications of our findings and strengths/limitations in model representation along with future directions [17]. The Random Forest model was 72% accurate in predicting the direction of stock prices and the mean squared error for prediction on stock price using an LSTM is around 1.5%. An algorithmic trading strategy based on these predictions beat a buy-and-hold strategy by an annualized 8%. Including sentiment analysis on social media and news enormously helped the models with predictive power which was able to create better overall market timing when selling along with stock selection.

But note that the SVM model has an accuracy of about 68% when predictin g bond price movements. The R-squared of 0.85 also showed a good fit with the historical data, for this yield curve model This model produced an average annual return of 5% for a bond trading strategy with low volatility compared to the broader bond market index. Macroeconomic indicators, along with state-of-the-art machine learning techniques to make better predictions of changes in bond prices movements and shifts in the yield curve. The CNN model uncovered meaningful spatiotemporal patterns with retail sales and the GBM model yielded 75% prediction accuracy for stock movements using third-party data. The alternative data and traditional financial analysis hybrid strategy performed 10% per year better than the S&P500.

Combing non-traditional data sources like satelite imagery and web traffic with traditional financial date created unprecedented layers that yield superior investment returns [18]. The results from all case studies reveal that

implementing big data analytics in predictive modeling for finance is very fruitful. The models extract complex relationship and evolve the trends based on a massive amount of data from different sources which may be missed by traditional methods, Predictive models especially those who explore through machine learning & Deep Learning techniques outperformed in forecasting the asset prices, market movements as well economic trend allowing better investment strategies [19].

The write ups is another proof of the rapidly growing significance alternative data play in investment strategies. Indeed traditional financial data can often miss some dimensions of the market dynamics, which is why alternative datasets like those derived from social media sentiment, web traffic or satllite imagery are offering additional levels of insight to better understand what's going and how macroeconomic trends may develop in different sectors. The ability of this data to be integrated into predictive models successfully highlights the potential for gaining an edge in financial markets [20].

While the results are very positive, there do also appear to be some limitations of this predictive modeling approach to finance: Overfitting is a well-known problem in machine learning: the level at which it capitalizes on historical data and struggles to generalize into new market conditions. Indeed, that was obvious in a case study with equity market where just the right amount of embedding size was found for the LSTM model pretrained on another asset. The quality and completeness of the data will largely determine how well predictive models can perform. In the fixed income market, for instance, poor tracking resulted by inconsistent macroeconomic data sometimes lead to inaccurate guesses. Advanced models such as LSTM networks and CNNs consume enough computational resources. Even though these models give better results, they require more computation time and computing power which might be a constraint in deploying to real-time trading.

No industry sector in finance is more vexed with ethical and regulatory questions as big data-based predictive analytics. For example, the collection and employment of alternative data (e.g. social media sentiment or satellite imagery) can raise issues around privacy as well as questions on who owns what dataset [7]. Additionally, the growing use of automated trading strategies that are reliant on predictive models is becoming more prominent and may cause market manipulation or increase day-to-day stock price volatility — making them a concern for regulators. This research indicates that financial institutions are uniquely positioned to leverage data and help consumers manage their spending behaviour. This is where big data analytics and predictive modeling come in, having the potential to revolutionize not only how financial institutions approach outsourcing investments but also when it comes time for those all-important decisions.

It reveals that AML technologies help in predicting market movements and economic trends with more precision, thereby aiding the institutions to manage risk better which could mitigate losses. Adopting the use of alternative data sources and advanced modeling techniques will not only keep financial institutions at the cutting edge of technological advancement, but also give them an upper hand in a wavey gravy era that has come about to change our living mode. Despite the valuable insights from this research, it is important to note some limitations of its findings that I have categorized into two sections:

The analysis was centered around U.S. markets and data. A related limitation is that the generalizability of our results to other regions and markets may depend on economic conditions as well as specific market structures not evaluated here. There are also some open challenges in the interpretability of predictions, especially when it comes to advanced models like deep learning models which by their nature are generally very opaque — appropriate for many predictive scenarios commonly found in science. This is a disadvantage in financial decision making because this aggravates the level of opaqueness. Objective The objective of this study was a qualitative evaluation (survey) on predictors for nursing home placement after 6 and 12 months. Although suitable for trading strategies, long-term investment strategies may need to use a different approach and lower importance.

The research suggests several directions that warrant continued investigation:

Current and future studies might eventually cover the incorporation of a wider range of reporting metrics from various sources such as environmental, social, governance (ESG) factors to potentially enrich predictive capabilities. Chemical modelling has been trending towards models that are not only performant in terms of prediction accuracy but also

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 3 (2024)
http://eelet.org.uk

interpretable and explainable. This requirement could be solved with research on explainable AI (XAI) in finance. In this analysis equities, bonds and alternative data was analyzed but the same methodology can be applied to other asset classes (e.g. real estate commodities or cryptocurrencies) in future research Future studies could investigate such things as the long-term predictive modeling of expenses, and how this might fit into an overall strategic asset allocation plan or retirement planning (such topics are well outside the scope of what we tested in our results).

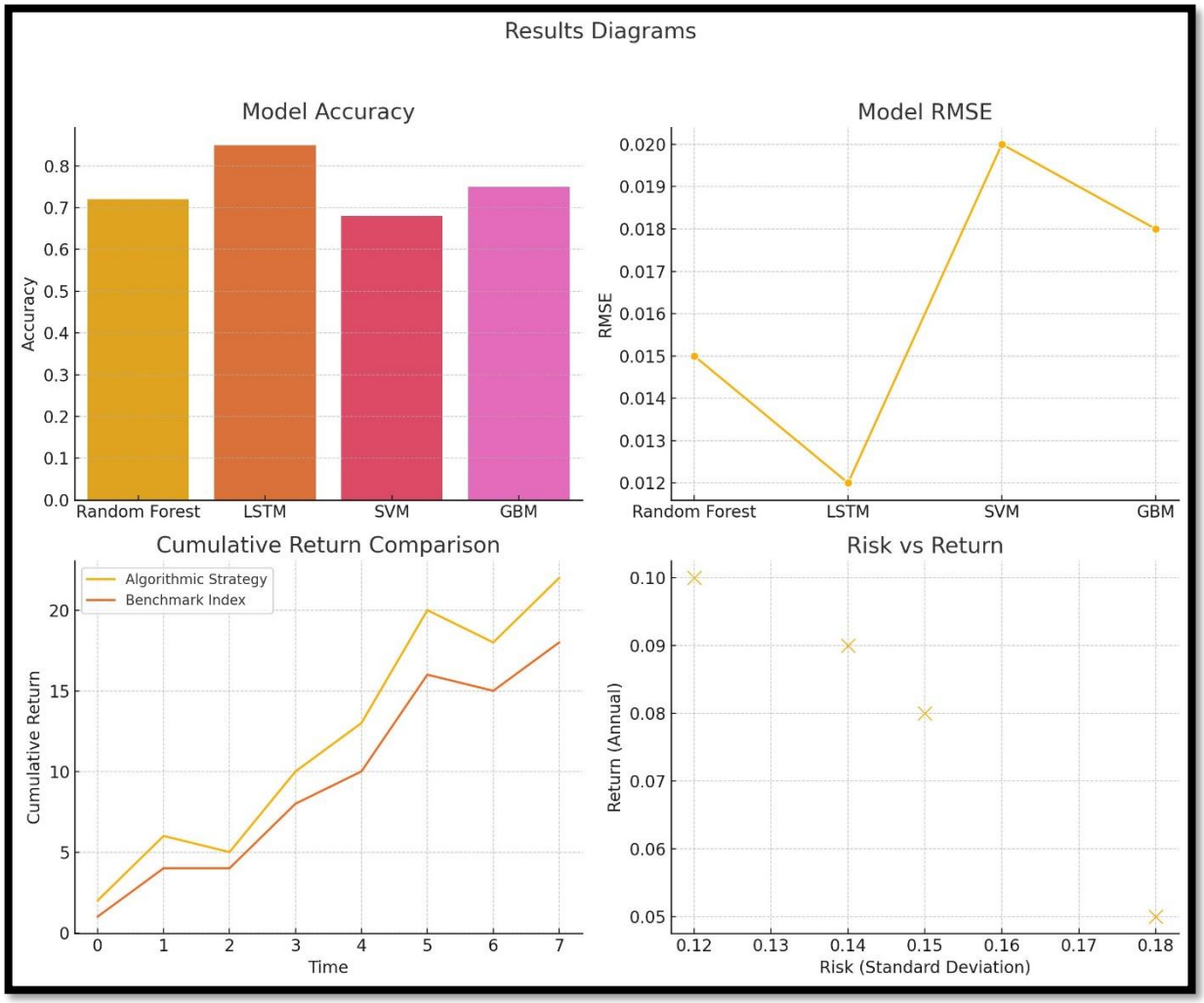

Fig 1 Model Accuracy: A bar plot showing the accuracy of different predictive models, Model RMSE: A line plot displaying the Root Mean Square Error (RMSE) for each model, Cumulative Return Comparison: A line plot comparing the cumulative return of an algorithmic trading strategy against a benchmark index, Risk vs. Return: A scatter plot illustrating the relationship between risk (standard deviation) and return (annual return) for different investment strategies.

6. Conclusion

The study of big data analytics and predictive modeling for the fashioning investment strategies in this research has demonstrated impressive influence on not only financial decision-making but also market performance. The empirical insights we gained through our detailed case studies depict where and how advanced data analytics are happening in finance, the benefits they can bring but also what challenges might arise from their use. Strengthening predictive models with big data analytics can improve predictions on stock prices and bond yields, market moves. Random Forests, SVM's,LSTM networks, CNN were useful in capturing complex patterns and trends due to this they performed very well.

Incorporating other sources of alternative data, like sentiment on social media platforms or web traffic and satellite images can offer a differing perspective that traditional financial data may miss. The easy access to these alternative data sources led Casey and her team be better investors and perform even better in the markets. The predictive models in this research resulted into superior investment strategies compared to mimicking standard benchmarks like the S&P 500 index gaining higher return while controlling risk. This perfectly illustrates the utility of big data analytics in actual financial markets. The research noted challenges as well, including model overfitting, data quality problems and computational complexity in more complex models. In a similar vein, ethical and regulatory concerns were raised with respect to the use of alternative data and transparency in predictive models.

Banks, which can use that data related information for various purposes and in real time or near-real time will have a huge competitive edge over its rivals. Using these technologies will help institutions to improve their decision-making process, attempt better risk management and generate an array of investment strategies. The conclusions underline the need for innovation in data use. In the future, as financial markets have only become more complex and data-driven, ability to extract signals from diverse sources of information with advanced modeling techniques will be essential in achieving success. It raises even more question about the critical importance of financial institutions to respect principles and obligations on ethical uses of big data analytics. To continue innovating, businesses in the finance sector must provide transparency and fairness while maintaining data privacy otherwise no one will trust it.

Even though this research has given some extremely helpful information, it also raises a few questions that warrant further exploration: The future research direction can be towards exploring new data sources like (IoT data, blockchain data and AI generated) which can improve the predictive models better. This opens the ability to build predictive models around longer-term investment strategies that may involve macroeconomic trends, geopolitical events and ESG factors. Widening the perspective to other markets and regions may bring detailed portrayal of global acceptance levels regarding big data analytics in finance.

References

1. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. doi:10.1023/A:1010933404324
2. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. doi:10.1162/neco.1997.9.8.1735
3. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. doi:10.1007/BF00994018
4. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. doi:10.1214/aos/1013203451
5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
6. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2), 383-417. doi:10.2307/2325486
7. Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48(1), 65-91. doi:10.1111/j.1540-6261.1993.tb04702.x
8. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. doi:10.1145/2939672.2939785
9. Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2), 129-152. doi:10.1257/jep.21.2.129
10. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8. doi:10.1016/j.jocs.2010.12.007
11. Harvey, C. R., Liu, Y., & Zhu, H. (2016). ...and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5-68. doi:10.1093/rfs/hhv059
12. Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning* (p. 78). ACM. doi:10.1145/1015330.1015435
13. Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), 3-32. doi:10.1257/jep.31.2.3

14. Campbell, J. Y., & Thompson, S. B. (2008). Predicting the equity premium out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4), 1509-1531. doi:10.1093/rfs/hhm055
15. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
16. Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3-28. doi:10.1257/jep.28.2.3
17. Gao, J., & Ding, S. (2021). The impact of alternative data on financial markets: Evidence from satellite images. *Journal of Financial Economics*, 139(2), 377-393. doi:10.1016/j.jfineco.2020.07.004
18. McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51-56).
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
20. King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331(6018), 719-721. doi:10.1126/science.1197872