ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

Comparative Analysis Of Conventional And Machine Learning Based Forecasting Of Sales In Selected Industries

***Samrat Ray**
PHD, Economics,
The International Institute of Industrial Management, Economics and Trade,
Peter the Great Saint Petersburg Polytechnic University, Russia
samratray@rocketmail.com
***Rohini Nikam**
PHD (Pursuing)Human Resource Management
SaiBalaji International Institute of Management Sciences,
Savitribai Phule Pune University, India
***Chhaya Vanjare**
PHD (Pursuing), Finance & Marketing
COL, DR. D. Y . Patil Vidyapeeth, Pimpri Pune
***Amruta Mandar Khedkar**
PHD, Business Management,
Modern Institute of Business studies Nigdi Pune,
Savitribai Phule Pune University, India

ABSTRACT

Sales Forecasting forms the heart of effective data driven decision making and affects every function in a business. There has been a quantum leap in technology choices available for predicting sales. Novel methods in deep learning and SVM have evolved while conventional time-series models like ARIMA, ARCH and Holts-Winter continue to be leveraged. Extensive progress has also been made on aspects such as explain ability and outlier treatment. There is however a gap in applying technology to business areas such as industries. Each industry is unique in terms of nuances and challenges it faces. These impact the forecasting process and hence should impact technology choices as well. This paper addresses that gap by proposing a novel recommendation framework comprising algorithms, accuracy metrics and seasonality treatment that have highest chance of successful application for respective industry. Since the top 3 industries that leverage Forecasting are BFSI (Banking, Financial Services, and Insurance), Pharmaceutical and Retail, we look at a cross- section between them. This paper analyses differences in forecasting through historic research, interviews with industry professionals and 3 experiments.

Some findings include Pharmaceutical being more seasonality driven whereas Retail being more causality driven owing to high impact of pricing and promotions. One logical conclusion is that conventional time-series models would have higher suitability with pharmaceutical while regression or neural networks maybe better fit for retail. Such thumb rules would aid easier and faster navigation in sales forecasting across industries.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

1. Introduction

Sales Forecasting forms a critical process in any business. Whether it is marketing, operations, human resources or finance, every function is impacted by it. Some examples include: Pricing and marketing spend planning is part of the sales forecasting exercise; Future plant capacity planning happens basis the targets; Hiring as well as appraisals are impacted by the same; There is no denying that accurate sales forecasting becomes the base layer of the building we call 'business'. It is interesting how the sales process and causal factors differ across industries. [2] and [3] favor the argument – 'Sales prediction is rather a regression problem than a time series problem'. However, in a pharmaceutical world, exogenous variables like price and marketing spends may not be important, while seasonality, especially weather seasonality may be critical for some medicines. On the other hand, insurance sales may be far more dependent on events like interest rate changes and budget announcement and may not have as much role of the time-series aspect. While Holts Winter may work well in Pharmaceutical, Regression/LSTM may better in insurance sales prediction – vice versa may not be true. As mentioned in [5] 'No single approach works best in every condition' – but can there be differences by industries with certain models performing mostly better in a certain industry? Different industries consume sales forecasting process in different ways since every industry comes with its own nuances, challenges, and unique needs. For example: Type of seasonality (Weather Vs Calendar): Pharmaceutical is mostly weather driven whereas Retail is mostly event and hence calendar driven. Finance does not seem as seasonal. Role of events/exogenous variables: Marketing spends, and promotions can never impact pharmaceutical as it impacts retail. Similarly, the launch impacts or loss of patents are more critical events in pharmaceutical than in retail. Calendar end, interest rate changes, budget announcements seem to be more important in the world of BFSI, i.e., Banking, Finance, and Insurance. Granularity of forecast is generally weekly or daily but may be a larger period in pharmaceutical. In the financial world over forecasting may not be as much of a challenge since the product doesn't expire, however in pharmaceutical true positives and false negatives both need to be balanced. Expired products are of far higher concern in Retail and Pharmaceutical but not in BFSI. Number of brands and SKUs can be enormous in retail, running into thousands and even millions Outlier treatment can be equally tricky, especially since a promotion-based peak may not mean permanent increase in demand but shifting of demand. There seems to be no presence of such comparisons and their 'SO WHAT', i.e., implications in machine learning. That is the gap we try to address here. The framework we would adopt in present in figure 1:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

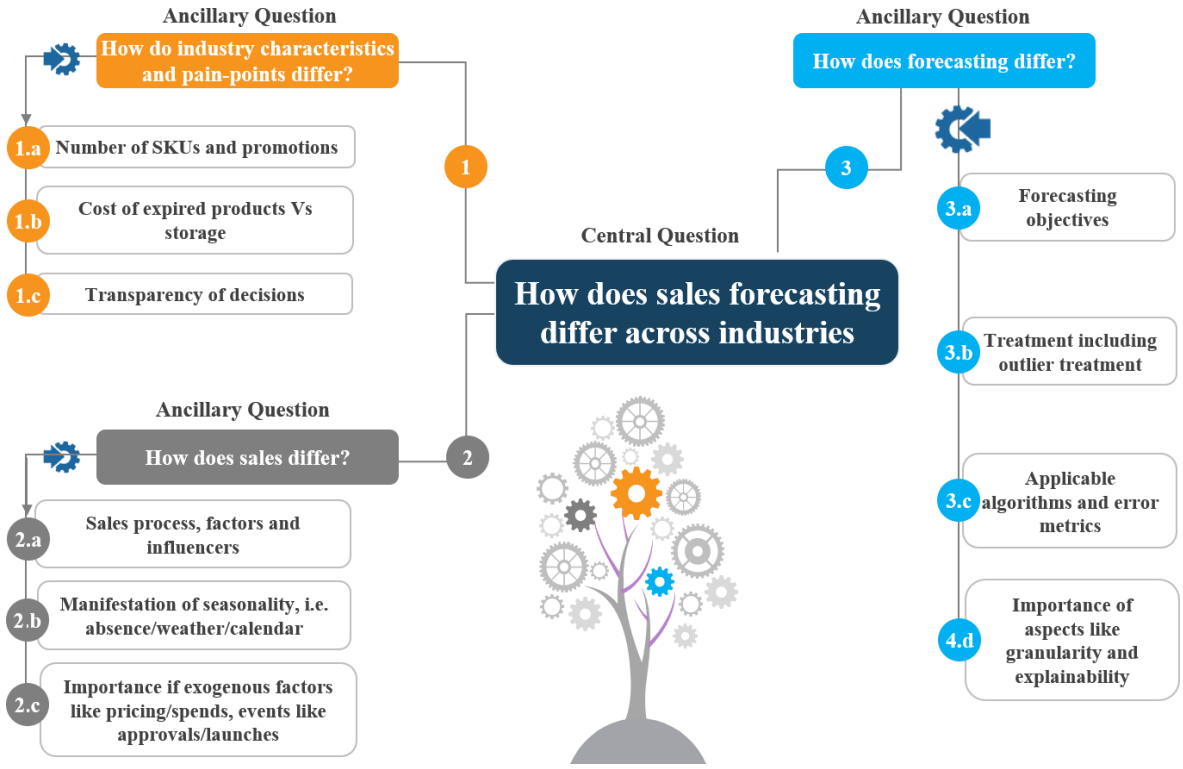


Fig. 1. Analysis framework for sales forecasting differences across industries.

In a nutshell, the 3 ancillary questions we are trying to answer while addressing the gap are (i) How do industry characteristics and pain-points differ? (ii) How does sales differ and (iii) How does forecasting differ? We further break each of these in sub-areas that this paper analyses on.

While a lot of research has been done over the past decade on usage of conventional time-series models such as Holt-Winters, ARIMA, SARIMA, SARIMAX, ARCH, GARCH on one hand, empirically novel methods in deep learning, SVM, and various combinations have also evolved. On the other hand, several accuracy metrics such as RMSE, MAPE, MAE and Bias are utilized to measure the accuracy. Recent advancements have also focused on explainable AI owing to adoption of deep learning and other black box methods.

However, the integration between technology choices and industry, for example:

-Does one algorithm or accuracy metric work better for certain industry?
-Does importance of time-series or causality differ?
-Does granularity of forecast or outlier treatment differ? This is the gap that is going to be addressed.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

1.1. How the recommendation will address industry differences:

This paper would provide a recommendation framework on sales forecasting differences in 3 largest industries that consume it. i.e., retail, pharmaceutical and BFSI. Our focus is on:

-Manifestation of time-series seasonality and key events product or competitor launches

-Impact of exogenous variables like pricing and marketing spends

-Machine learning differences in form of algorithms, time granularity(weekly/monthly), type of accuracy measure etc.

-Another aspect could be the granularity of forecast(daily/weekly/monthly), for retail even a day can vary hugely, and hence a weekly granularity is essential. Pharmaceutical may work well even at a month level

-Outlier treatment and related challenges

- Metric: Should accuracy be as important in financial products (don't expire) Vs Pharmaceutical. Should we select Accuracy/Recall/Precision? Should we adopt MAPE/R3M/RMSE etc.

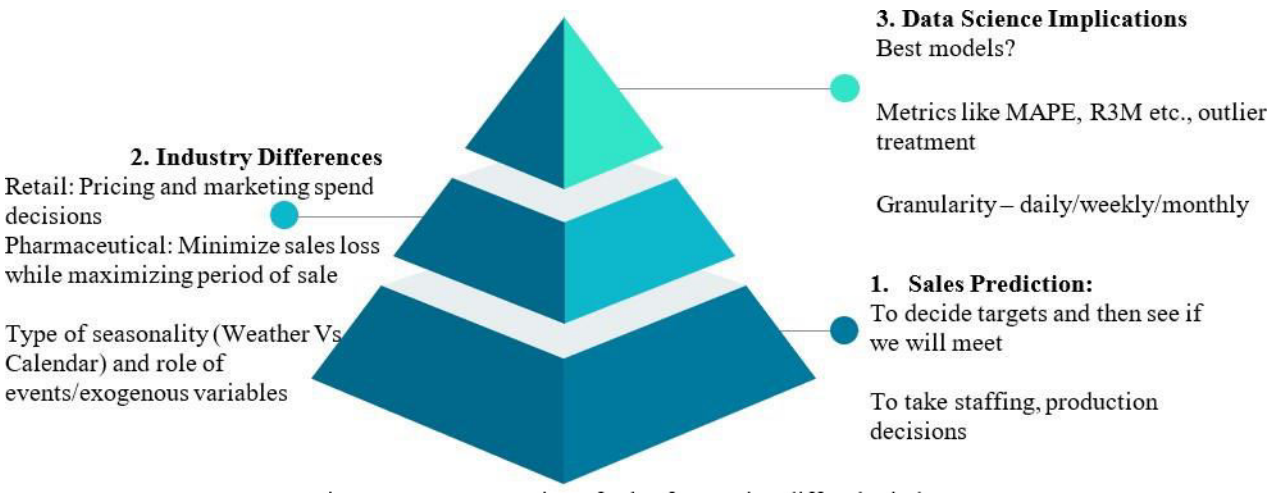


Fig. 2. How consumption of sales forecasting differs by industry

To summarize the objective, each industry conducts sales forecasting to cater to common decision making. There are some additional reasons or decisions for specific industries which impact the kind of models and methodologies adopted. The rest of this paper is organized into 7 sections. Section 2 looks at related work followed by statistical definitions and preliminaries in section 3. We deep-dive into methodology in section 4 and then section 5

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

leads to experimental evaluation and results analysis. Discussion of findings and their implications form part of section 6, post which we conclude with a novel recommendation framework in section 7.

2. RELATED WORKS

This section studies the evolution of technology with regards to sales forecasting and possible linkages with industries.

2.1 Ensemble techniques and Neural networks Results and Discussion

Extensive enhancements have occurred technically. B2B Sales forecasting is far more complicated, and hence we may look at B2B scenarios separately or explicitly focus on B2C scenarios as mentioned in [1].
Recent research [1][2][5][9][13][14, 15] suggests that Ensemble methods like stacking, combination methods and Neural networks may give better results (i) For retail scenarios - a stacking approach including more regression models' gave better accuracy. In [13], combination techniques gave statistically significant increases in forecasting accuracy. (ii) hybrid model comprising ARIMA+ANN as well as combination of other neural networks and conventional techniques as indicated in [5]. (iii) ANN (p, d, q) modeling helps with more precise time series forecasting. (iv) In [14, 15], authors use a combination of ARIMA and LSTM have higher accuracy in some cases, while Winters and ARIMA models are most consistent. (v) In [18] for insurance sales forecast, other models were outperformed by the improved LSTM.Using machine learning algorithms to predict academic activities, such as academic accomplishment and student learning styles, gives a more accurate approach of foreseeing academic events. In [27] the education industry, machine learning algorithms including K-nearest neighbour (KNN), random forest, bagging, artificial neural network (ANN), and Bayesian neural network (BNN) are being utilised to forecast future events.Machine learning algorithms are used by educational institutions to especially relevant' instructional practices and educational attainment, allowing them to identify at-risk pupils early and design strategies to help them overcome their restrictions. Given the advantages of machine learning, there are still some disadvantages which might limit its accuracy or application in anticipating academic outcomes, also including mistake predisposition, data collection, and moment issues.

2.2. The rise of explainable AI

New explainability methods for block-box or ensemble techniques can be tried other than modern age explainable AI like LIME/SHAP. Better models may be more explainable as mentioned in [1]. (i) In [1], author speaks of EXPLAIN method, which uses sensitivity analysis, i.e., determining variable important by removing or changing variable value (ii) The IME technique also considers inherent relations between 2 input variables.

2.2.1 Causal Vs time-series weight age shift between industries

A lot of research has been done on sales forecasting as an area, like in [2], [26] and [27] conventional models like, GARCH, ARIMA, SARIMA, SARIMAX, Holt-Winters as well as combination of forecasts produced by different algorithms is mentioned. Prophet is widely

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

used as a tool. In [13], ETS and ANFIS have been introduced as well. (i) Others: Linear models, Logistic, Tree based like Random Forest, Bayesian (ii) New areas including Deep learning [3] and SVM [4] have been tried. (iii) While ANN have mostly outshined over linear methods, Box Jenkins performs similarly in case of long memory; Exponential smoothing performed better for yearly projections and similar for quarterly ones. ANN may not pick up seasonality well [5]. (iv) In [8], deseasonalizing of the sales data was not required by time lagged FFNN's. This indicates that neural networks may correctly detect and utilize seasonality during both training and Forecasting.

2.2.3 Metric for model accuracy

Metric for model accuracy may differ: MAE Vs MAPE Vs RMSE; Accuracy Vs Precision Vs Recall: (i) In a pharmaceutical scenario, both true negatives and false positive may lead loss of revenue/profit, which may not be so in a financial scenario like insurance, because the products will not expire. (ii) A common challenge that emerges across [1], [2], [5] as well as in interviews is the challenge of 'Small data problem', having insufficient history to predict. (iii) Few papers such as [7] have applied learnings from sentiments unstructured data and K-means in [23-37] to increase the quality of output

2.3 Findings from discussion with Industry professionals

Financial service industry continues to use statistical and regression-based models: (i) Logistic regression was used to forecast in a credit card default problem. Neural networks along with explainable AI to understand the accuracy loss. The cadence was monthly/quarterly and average accuracy was ~80%, (ii) In another B2B Deposit problem statement, statistical methods like Lasso, Ridge, Error correction and SORE were used. Macro-economic factors like GDP, profits, exports imports, inflation were key inputs. Accuracy was generally <50%. Scenario planning was key.

Pharmaceutical does benefit from conventional techniques: (i) For a top 10 pharmaceutical giant, combination of various time-series techniques like SARIMA (Seasonal ARIMA) and Holts-Winter and then synced mathematically with other variables such as market demand
Retail requires a high level of creative problem solving before applying machine learning: (i) For an apparel client, a combination of LTSM, XG-Boost and causal techniques was used and average accuracy was ~75%. However, the key challenges included (ii) Outlier treatment is critical and can't always be done statistically. This also makes history unreliable for forecasting (iii) Factors such as warehouse cost need to be input to the decision-making process (iv) For another FMCG based retail set-up, 1000s of models were created for 70-80K SKUs and forecasting was carried out via univariate techniques like ARIMA/LSTM/LRS and Prophet. The overall process required applying a series of business rules and promotional/event impact followed by forecasting. (v) Outliers a definite challenge, for e.g., stocking due to a promotion may be picked by models but is not expected to repeat. Every month, SKUs to be run manually Vs automatically are chosen. MAPE/Bias/error/variability/market shares are various metrics that are considered. SKUs may phase in and out and such combinations need to be identified and their history needs to be combined.

2.4 STATISTICAL TERMS

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

- Time-series: Time-series data refers to data that is organized by dates at regular intervals like daily, weekly, monthly, quarterly, or yearly. Time-series data may also be auto-regressive since it holds high correlation within different periods. Key components include trend, seasonality, and irregularity. A rarer component is cyclicity.

-Causal: Data comprising variable that have cause-effect relations are causal in nature. An example may include impact of price of supply. Causal data may not have regular date values but do require data of all input variables.

-Seasonality: The period at which the patterns in time-series data repeat is called seasonality. Generally, 1 year, i.e., 52 weeks or 12 months or 4 quarters depict one seasonal period. Can be weather driven or calendar driven

-Accuracy metrics: In forecasting, various metrics like MAPE (Mean Absolute percentage error),MAE (Mean absolute error), RMSE (Root mean squared error) exist. Some organizations also use novel metrics like R3M (Rolling 3 months) to reduce the penalization due to weather changes.

-Neural networks: A section of machine learning called deep learning involves using neural networks. They mimic the way neurons send signals in the human brain and hence are very complex and black boxes. They generally give higher accuracy. Few types include RNN (Recurrent Neural network) and LSTM (Long short-term memory).

-Explainability: Ability to explain the output of a machine learning model and link it back to the input variables is called explainability. The cause-effect understanding is vital in many decision-making processes. The success as well as inability to explain neural networks has led to the invention of explainable AI like LIME and SHAP.

-EDA or Exploratory data analysis: EDA refers to the process of understanding variables, their distributions, and relation with each other through visual and numeric means. Some common aspects include 5-point summary, distribution plots, pair plots, correlation plots and pivots. EDA generally guides the next course of action.

2.5. METHODOLOGY

During the course of this paper:

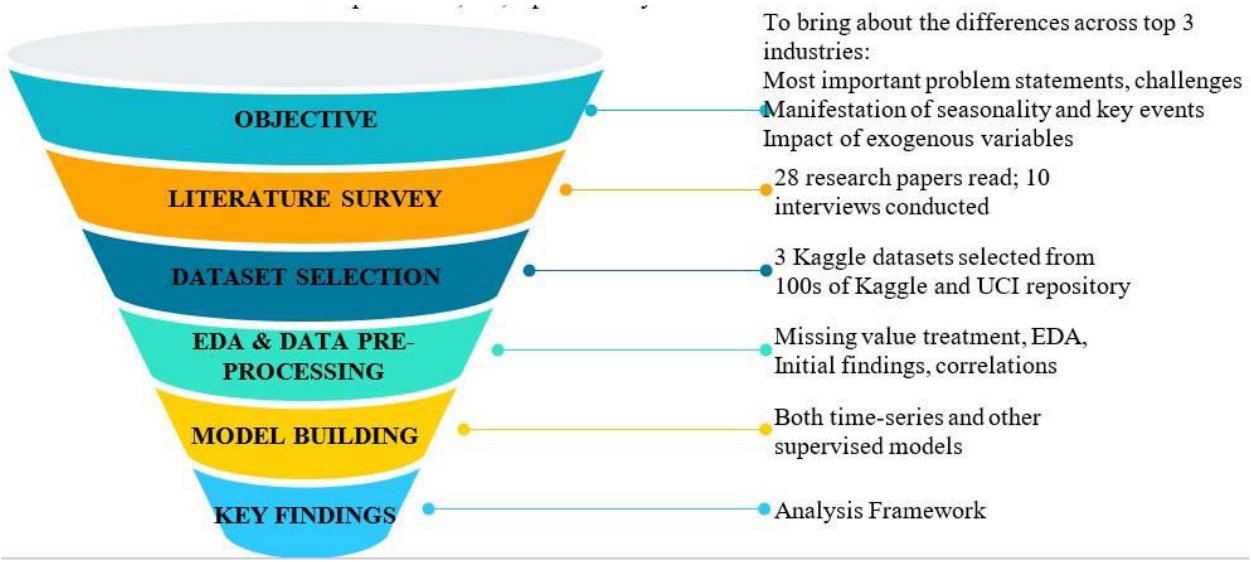
- 28 research paper were studied

- 10 interviews with industry professionals were taken

- 3 cases were implemented.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022



The key stages in the paper include objective finalization, literature survey, dataset finalization for 3 experiments, analysis and model building and finally summarizing key findings.Primary analysis was conducted as an input for the recommendation framework. 3 sales forecasting datasets, 1 from each industry were selected to forecast.

2.5.1 Software and Hardware Requirements

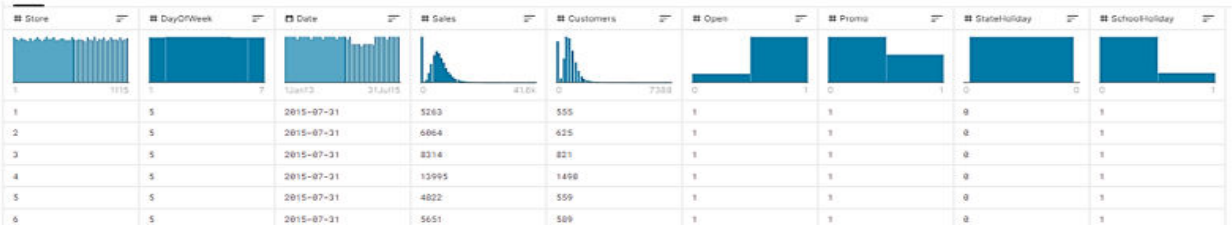
The experiments were performed using Google Colab and hence special software and hardware aspects were not required.

2.5.2 Dataset Details

3 datasets from Kaggle have been used to analyze and apply sales Forecasting across industries. This section helps deep-dive into the simulation setup datasets, exploratory data analysis, as well as model performance and next steps.

3. Analysis

3.1 References Retail: Rossmann Store Sales



About: The ask is to provide 6 weeks of daily sales forecast for 1,115 of the 3,000 stores Rossmann has across 7 European countries. Several factors such as promotions, holidays distance from closest competitor store influence store sales. The holidays could be state and school holidays and there may be underlying seasonality at play. 2 datasets are provided, their features and screenshots follow:

Sales: Number of Store, Count of DayofWeek and Date, Number of Sales and Customers, Flags for Open, Promotion, Sate and School holidays Number of rows = 1,115, number of columns = 10

Store: Number of Store, Number of Storetype, Assortment, Competition distance and distance for closest store, promotion details Number of rows = 1.02m, number of columns = 9

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

Figure 4: Overview of Rossmann store sales dataset

We observe the following: (i) There are 9 columns in each of the datasets which may need to be joined for further analysis (ii) Dataset is a mixture of time factors such as dates, flags such as holiday and other dimensions. We further deep-dive in next section:

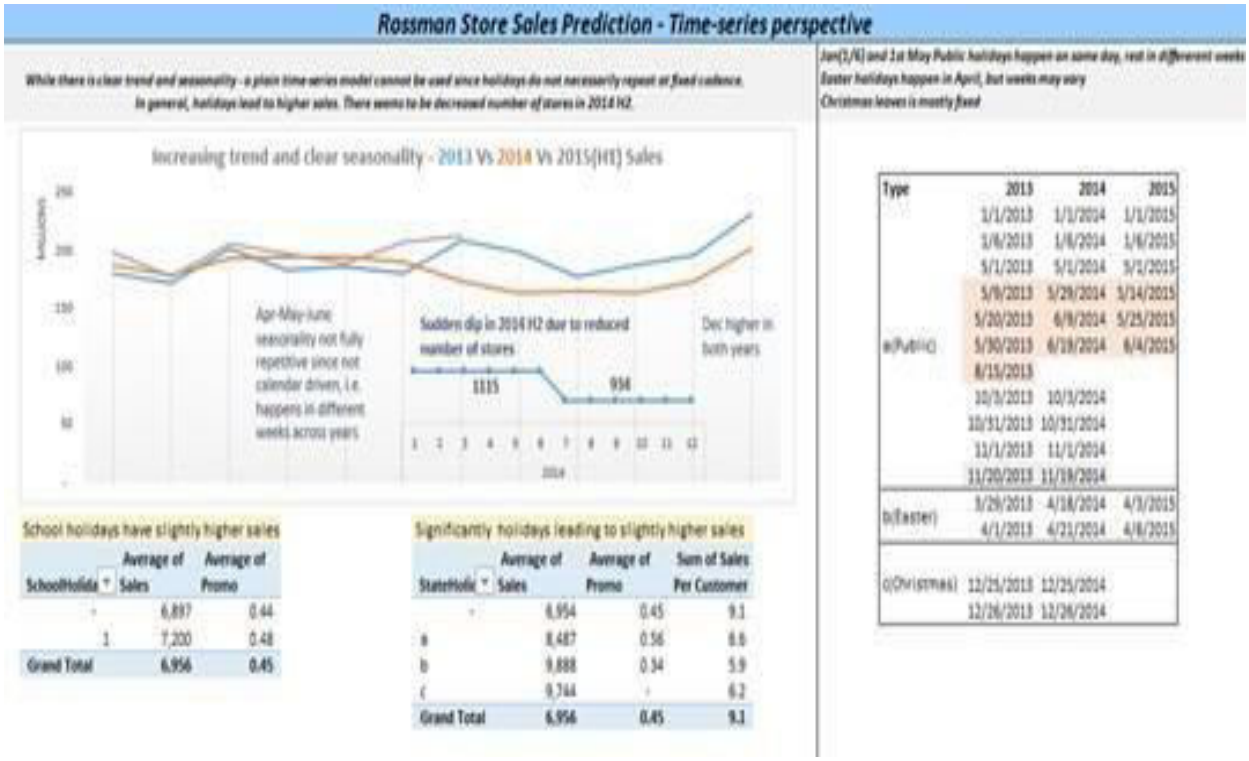


Figure 5: Rossmann stores sales exploratory data analysis from time-series and causal perspective

3.2 Exploratory data analysis: The above figure is a representation of some visual analysisperformed answering questions such as:

- What is the impact of holidays on sale?
- Does seasonality exist?
- What is the trend?

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

3.3 Findings from EDA:

- While there is clear trend and seasonality - a plain time-series model cannot be used since holidays do not necessarily repeat at fixed cadence. In general, holidays lead to higher sales. There seems to be decreased number of stores in 2014 H2.
- Jan (1/6) and 1st May Public holidays happen on same day, rest in different weeks
- Easter holidays happen in April, but weeks may vary
- Christmas leaves is mostly fixed
- Causal findings:

1. Sales and customers have high correlation so we do not need to use both
2. Sales patterns by day, promotion and store follow
   - *17% datasets (172,818 of 1,017,212) records had 0 sales and were not open and henceremoved
- Competition not necessarily impacting stores

Methodology: The dataset showed both time-series aspects as well as cause-effect relationships. As a standard, an 80-20 train-testing was adopted. Various methods tried:

- Regression solutions like Linear, Ensemble methods, LightGBM (Light Gradient BoostingMachine)
- Metrics like MAPE, MPE, RMSE were checked. However, MAPE was adopted since it is apercentage and also does not suffer from sign cancellation problem

- Best MAPE for LightGBM was 8.86%

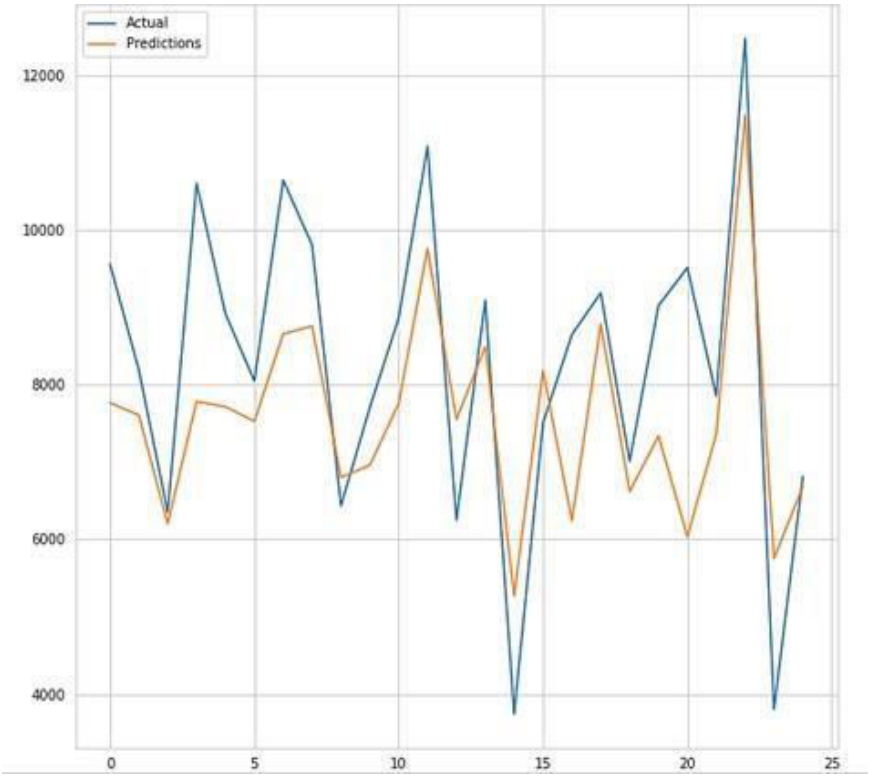


Figure 6: Rossmann stores sales actual vs predicted for one store

While prediction for stores differed, above is a representation of accuracy for one of them across time. Way ahead: The other possible seem solutions that can be tried for Rossman Store Sales:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

•SARIMAX (currently thew error), ARIMAX with Seasonality Index or holiday flags
•Hierarchical forecasting
•Clustering stores and then predicting Suggested Metrics: Weekly MAPE or RMSE

3.4 Pharmaceutical: Pharma sales data

About: The ask is to forecast monthly sales for 8 drug categories based on 6 years of history. The dataset is consolidated from 600K transactional data for 57 drugs. These are clubbed into 8 ATC (Anatomical Therapeutic Chemical) classes. ATC is controlled by the World Health Organization and is a classification of the drug active ingredients according to the organ or system on which they act. The data comprises hourly, daily, weekly and monthly sales for each of the 8 categories. Number of rows = 2,106, number of columns = 8
It is to be noted that datasets with good quality datasets with possible cause-effect relationsand time-series data were searched but not found for pharmaceutical industry. This further indicates how unique pharmaceutical industry is. In the below figures, we analyse

1. Distributions of various drug groups
2. Presence of seasonality in different groups
3. Whether seasonality is better at weekly or monthly levels

	Year	Monthly Correlation (Vs Next Year)	Weekly Correlation (Vs Next Year)
M01AB	2014	-56%	-11%
	2015	56%	34%
	2016	8%	22%
	2017	-16%	7%
	2018		
M01AE	2014	28%	9%
	2015	-25%	8%
	2016	-20%	31%
	2017	-21%	29%
	2018		
N02BA	2014	31%	31%
	2015	29%	48%
	2016	60%	55%
	2017	36%	32%
	2018		
N02BE	2014	76%	58%
	2015	81%	66%
	2016	83%	81%
	2017	71%	60%
	2018		

	Year	Monthly Correlation (Vs Next Year)	Weekly Correlation (Vs Next Year)
N05B	2014	27%	26%
	2015	-22%	-2%
	2016	-15%	9%
	2017	-42%	4%
	2018		
N05C	2014	39%	24%
	2015	47%	7%
	2016	29%	-5%
	2017	-52%	-13%
	2018		
R03	2014	65%	26%
	2015	69%	48%
	2016	41%	11%
	2017	64%	22%
	2018		
R06	2014	82%	58%
	2015	86%	68%
	2016	88%	70%
	2017	84%	65%
	2018		

Figure 7: Pharma sales dataset overview, yearly and weekly performance Findings from Exploratory data analysis

- Trends for most drugs not linear and not clearly increasing or decreasing across time. We should be expecting high error in output due to the same.
- Not all drugs are seasonal. The correlation between different years ranges from single digit or negative percentage to eighty percentage
- Additionally, since monthly correlations are higher than weekly, it may be better to predict monthly sales. This may be because pharmaceutical is more driven by weather driven seasonality which not as repetitive at a weekly level.

Next, we check distributions between drug categories and if there are any patterns that exist.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

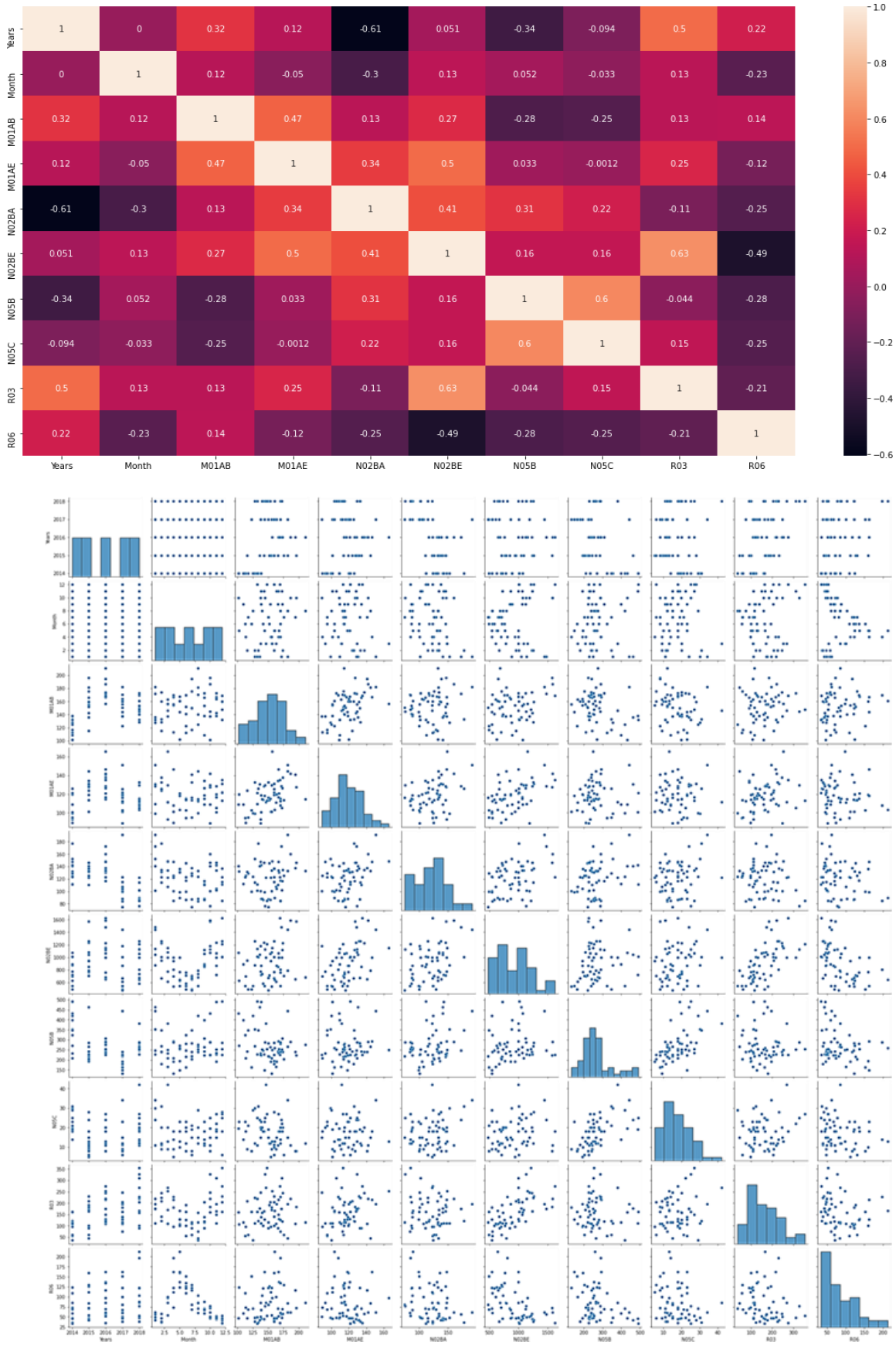


Figure 8: Pharma sales dataset variable pair plot showing various distributions and relations between variables

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

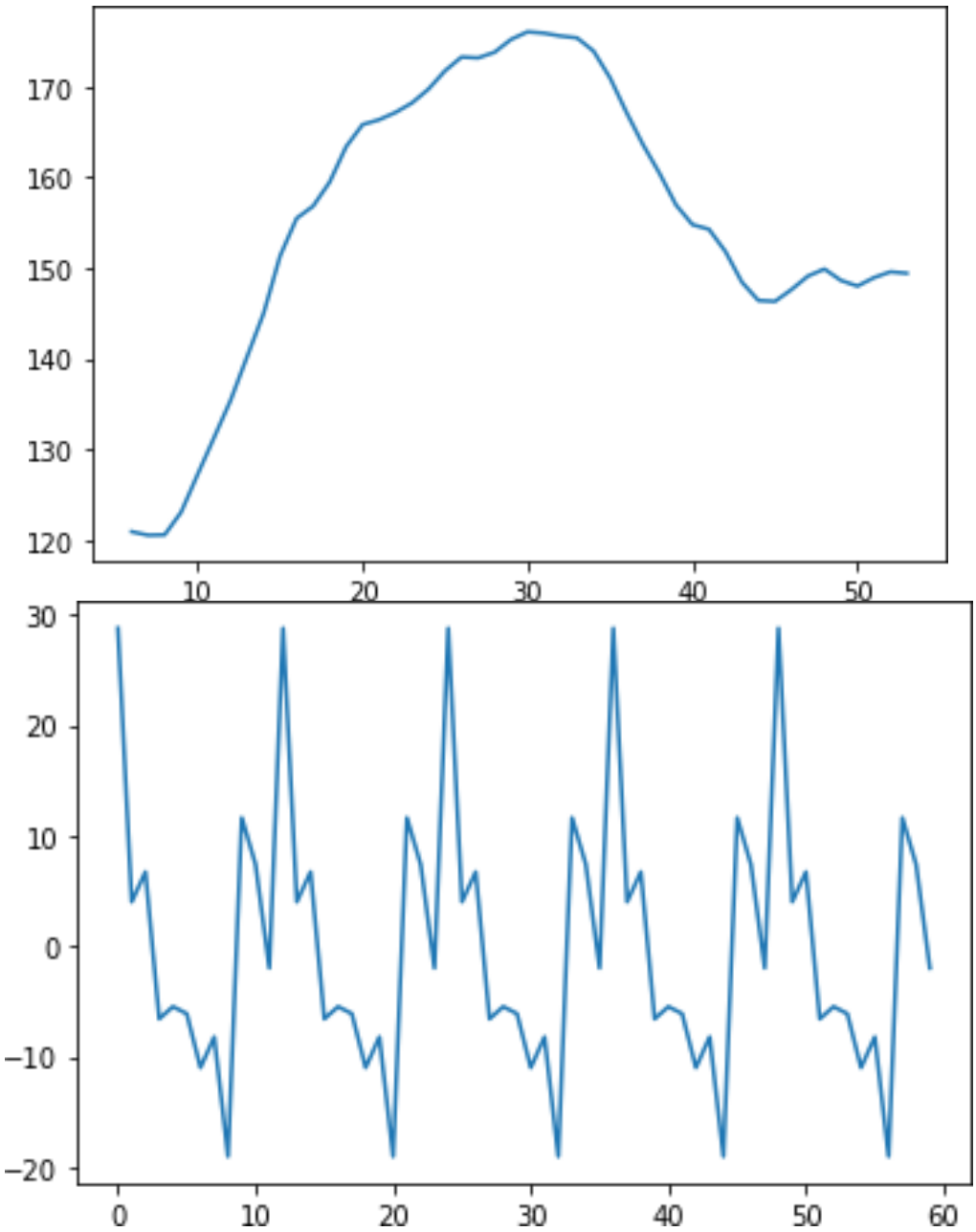
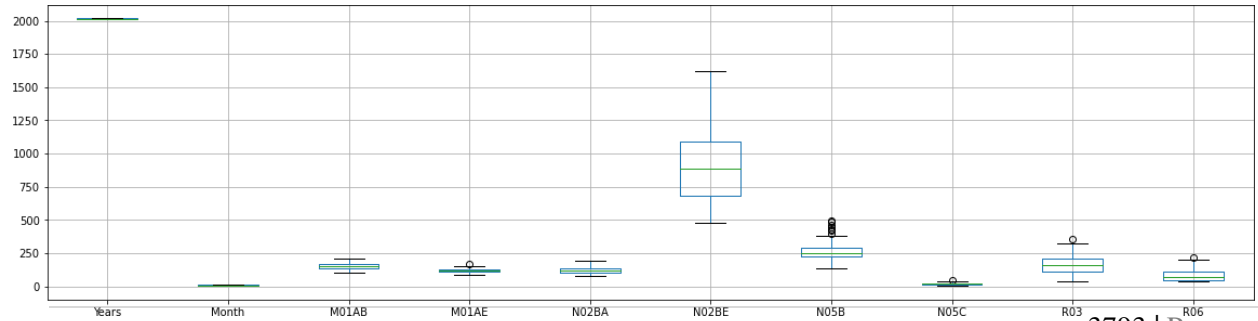


Figure 9: Pharma sales dataset variable correlation, trend and seasonality

- Most of the columns seems uniformly or normally distributed
- Outlier treatment is done by using median as replacement values. Since the data is skewed, mean is not a good choice for outlier treatment.



ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

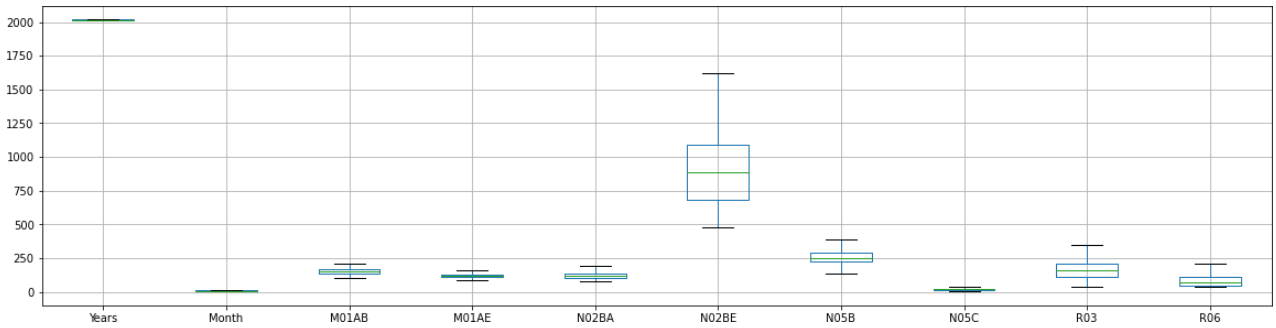


Figure 10: Pharma sales data outlier treatment – before and after M01AE and N05B show no outliers anymore after treatment

3.5 Methodology

Various conventional methods like Exponential Smoothing, Holts Winter, ARIMA were attempted. Since we have univariate data, regression-based methods would not be applicable. We do not have enough data for deep learning.

- MAPE was 7+%Sample MAPE:

1. M01AB 7.91
2. M01AE 7.154
3. N02BA 12.29
4. N02BE 20.05
5. N05B 26.40

Way Ahead: Other Possible solutions that can be tried for pharmaceutical drug sales:

- Further clustering for highly collinear drug sets maybe possible
- Trying other models

Suggested Metrics: Monthly MAPE or Rolling 3 months(R3M) owing to weather seasonality being at play, which is uncontrollable.

4.0 BFSI: DJIA 30 Stock Time Series

About: The ask is to predict stock market data for a particular stock basis 12 years of history. We have chosen the IBM stock dataset. The data is organized in following columns:

- Date, Count of Open rate Number of High in the day, Count of Low in the day, Count ofClose rate and Volume
- Number of rows = 3,020, number of columns = 7

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

Figure 11: IBM stock dataset overview, describing sample rows and distributions of all the columns

Exploratory Data Analysis: The questions to be answered through the analysis include 1.estimating relationship between close, open, high and low
2.Absence or presence of trend 3.Absence or presence of seasonalityThe findings from EDA include:

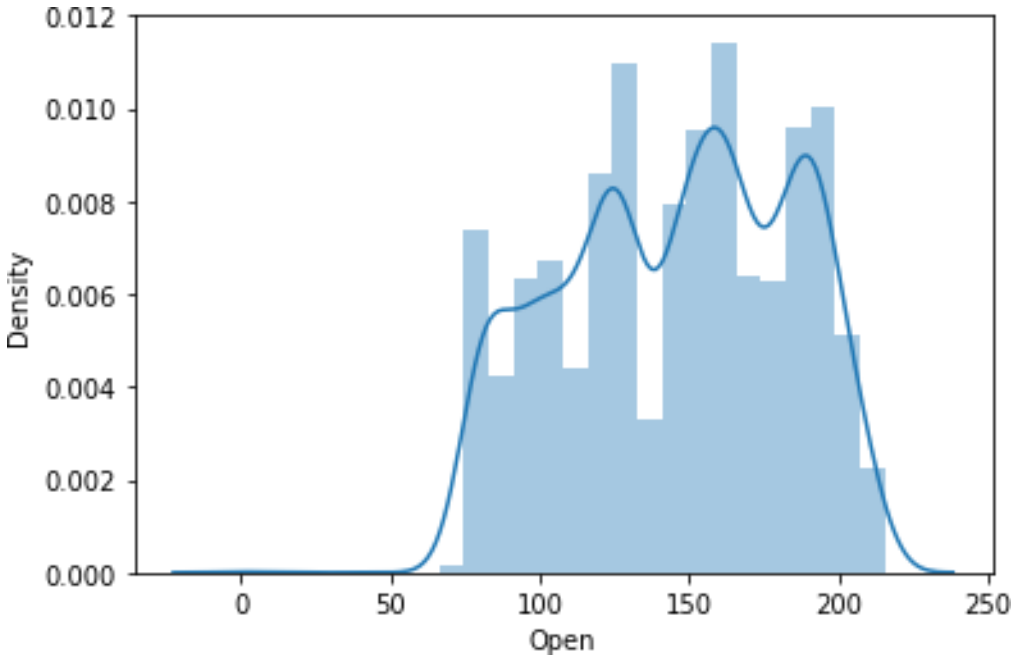
•The close, open, high, and low are closely correlated

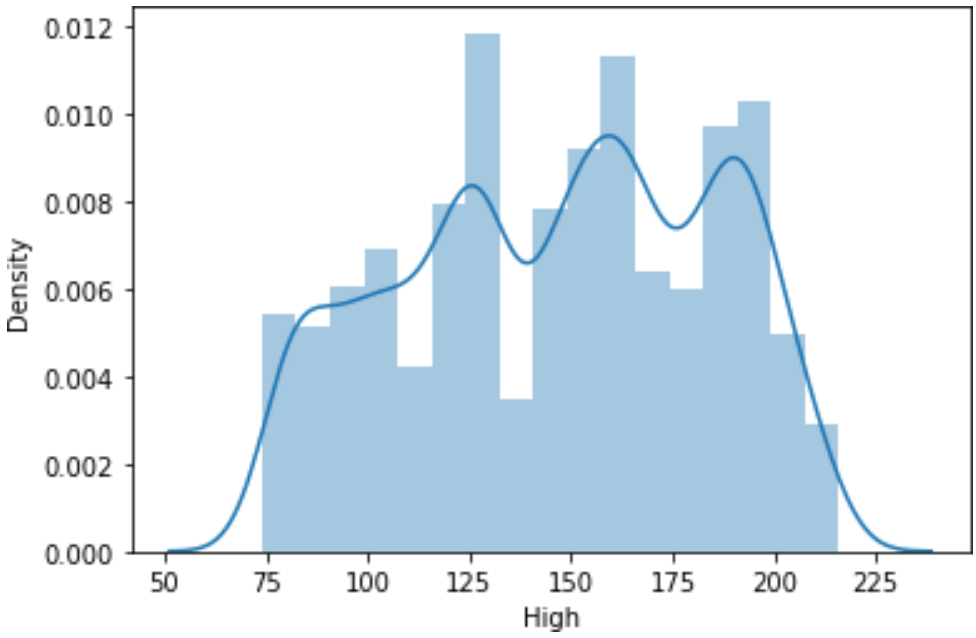
• Data does not seem to show seasonality or clear trends though we have 12 years of data, anda date is associated, hence solution will have to be focused on cause-effect relations.

Figure 12: IBM stock open/close and volume performance

ISSN PRINT 2319 1775 Online 2320 7876

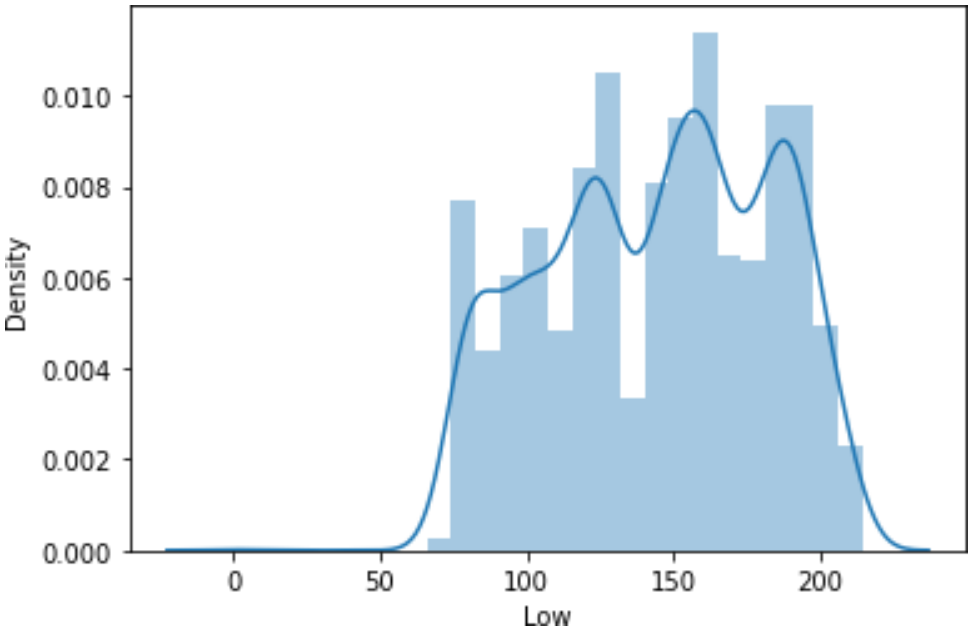
Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

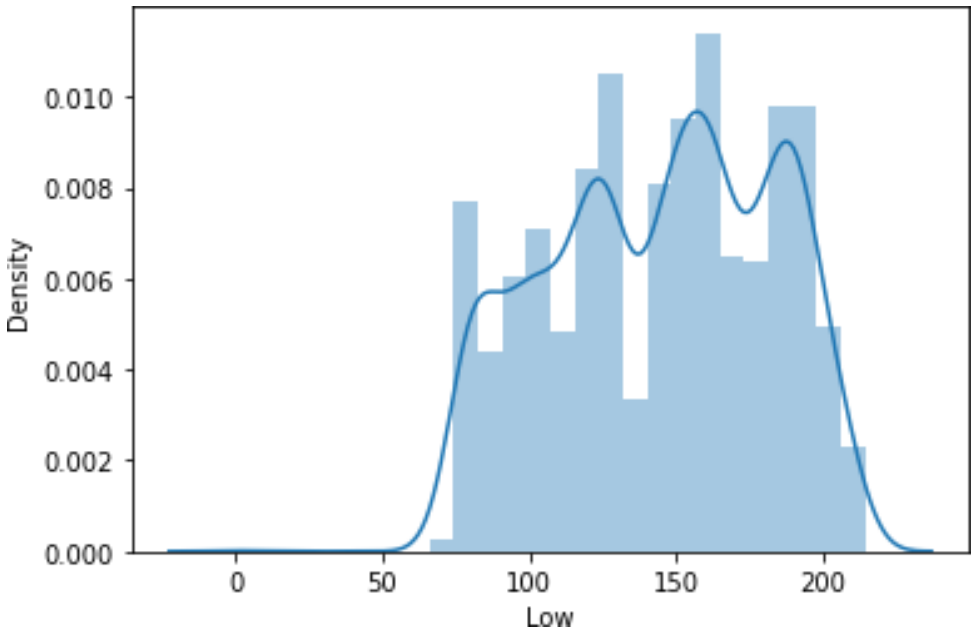




ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022





ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

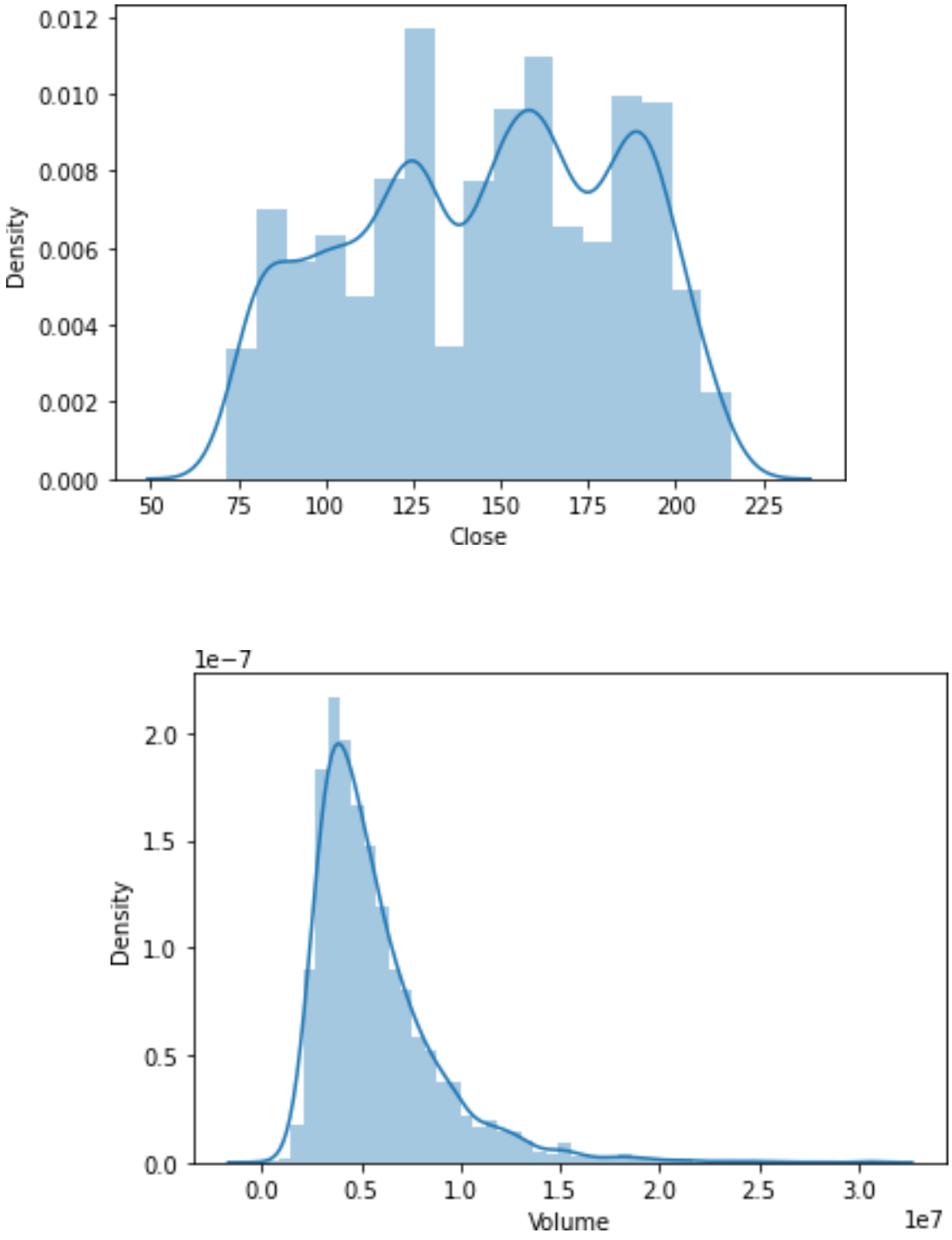


Figure 13: IBM stock variable distributions

4.2 Methodology:

• Regression models like Linear and Neural networks like LSTM tried to predict Volume andClose

•Best MAPEs: Volume 31% and Close 6%

A key question here is to judge whether the low Volume accuracy is acceptable. Basis interviews with industry ofessionals, financial accuracy is generally on the lower side since we can best use macro factors which are not strong

predictors but can't be ignored.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

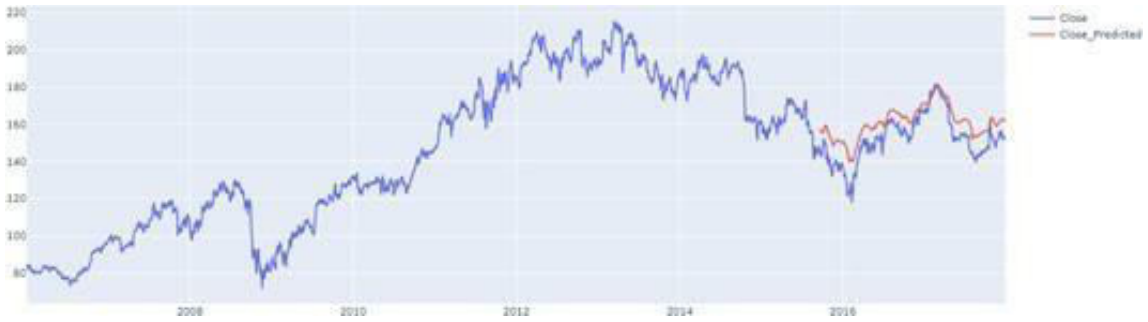


Figure 14: IBM stock close predicted vs actual

Way ahead: Other Possible solutions that can be tried for financial sales:

•Other models to be tried

• List of key financial events and other variables can be createdMetrics: Weekly MAPE

5. DISCUSSION OF RESULTS

The 3 prongs of research - historic research papers, interviews or the 3 experiments point to similar insights: (i) Each of the industries have a different weightage of causality- time-series importance, granularity of forecast as well as algorithmns that worked well. (ii) A key challenge is retail industry is the quantum of SKUs, often running in thousands or millions. Volume(rows) of forecast was highest in retail across discussions and datasets. Hence retail will need some creative problem solving or other methods like clustering even before specific algorithms are applied. (iii) Seasonality doesn't seem to impact the financial set ups as much, hence leaning to regression/statistical based solutions where explainability is important like B2C sector, and neural networks mostly for company internal problems. Explainable AI may be adopted to explain the difference between simpler ML models and neural network-based models. (iv) Retail forecast is generally more granular than pharmaceutical, generally done at daily or weekly levels (v) Customer behavior buying patterns do not seem as important in pharmaceutical side as in other 2 industries. Macro-economic factors seem to be important factor in financial forecast across the interviews, even though the overall model accuracy may not be as high (vi) Pharmaceutical world may be impacted most by weather driven seasonality, though not all drugs are seasonal. Since weather pattern changes every year, the resultant seasonality would also be impacted. (vii) Key events and methodology to calculate their impact is unique to every industry:

•Product launches, whether own or competitor launch are key across

•Macro-economic changes are key to financial space

•Loss of patent or clinical trials are most important events in Pharmaceutical

•Retail is filled with end-less promotions in form of advertisements and discounts

Basis all the research and analysis, we summarize our sales Forecasting framework across industries as follows:

Area	Description	Pharmaceutical	Retail	BFSI
Seasonality	Importance andtypesofseasonality	Highontime-series		HighonCausality (Cause Effectrelations)
		M/H	M	L
		Weathermostly	Calendar	Calendar

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

Events	ImpactofExogenousvariables	LOP,clinicaltrial	Severaltypesincludingspends,price, competitorentry	Macro-economicfactorslikeBudget, quarterend,Interestratechanges
Techniques	Suggestedalgorithms	conventionaltime-series modelsworkwell	SARIMAX,time-series,NNaswellasexpectedtoworkwell	Regression/statistical(externalclients)/ LSTM(internal)expectedtoworkwell
Granularity andAccuracy measure	Week/month	Monthisgoodenough R3M	Week MAPE	WeekorMonth MAPE
OutlierTreatment	Importanceandquantum	Low	High	Medium
Explainability	Explanationofresults	May/Not	May/Not	Important

6. Conclusion

While the exact solution framework would depend on factors more than industry – such as data available, B2B Vs B2C, size of company etc. – Industry does play an important role. Most findings from previous research, interview with industry veterans as well as the experiment narrow on some thumb rules. These recommendations relate the nuances, priorities and challenges the industry faces to technological translation in terms of algorithms, granularity of forecast, accuracy metrics etc. Key conclusions include:

• Pharmaceutical data is far more on time-series side while financial data is far more causal,and retail is generally a combination

• Forecast will be far more granular for Retail since decisions may be taken at daily level,whereas for Pharmaceutical even monthly could suffice

• Conventional time-series models would general work well in pharmaceutical set-up whereasregression and neural network methods would generally work well on financial side.

The next steps would be to:

•Test the recommendations on a larger sample to increase the confidence levels

•There is a possibility to extend the research to more industries and areas

• Check how forecasting methodologies have changed post events like the pandemic. Compare recommendations not just on accuracy but also consistency.

Declarations

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

The Author declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Conflicts of Interest

This section is compulsory. A competing interest exists when professional judgment concerning the validity of research is influenced by a secondary interest, such as financial gain. We require that our authors reveal any possible conflict of interest in their submitted manuscripts.If there is no conflict of interest, authors should state that "The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper."

References

[1] Aras, S., Deveci Kocakoç, İ., Polat, C., 2017. Comparative study on retail sales forecasting between single and combination methods. Journal of Business Economics and Management 18, 803–832.

[2] Bohanec, M., Borštnar, M.K., Robnik-Šikonja, M., 2017. Explaining machine learning models in salespredictions. Expert Systems with Applications 71, 416–428.

[3] Cohen, M.C., Gras, P.E., Pentecoste, A., Zhang, R., . Demand prediction in retail .

[4] Cracan, C., 2020. Retail sales forecasting using lstm and arima-lstm: A comparison with traditionaleconometric models and artificial neural networks .

[5] Dragomirescu, S.E., Solomon, D.C., 2010. A case study concerning sales prediction using sales quantitative prediction methods. The Annals of "Dunarea de Jos "University of Galati. Fascicle III,Electrotechnics, Electronics, Automatic Control, Informatics 33, 23–28.

[6] Fan, Z.P., Che, Y.J., Chen, Z.Y., 2017. Product sales forecasting using online reviews and historical sales data: A method combining the bass modeland sentiment analysis. Journal of business research 74,90–100.

[7] JIANG, Y.P., 2017. Prediction of the national total retail sales of consumer goods based on arimamodel. DEStech Transactions on Social Science, Education and Human Science .

[8] Kaneko, Y., Yada, K., 2016. A deep learning approach for the prediction of retail store sales, in: 2016IEEE 16th International conference on data mining workshops (ICDMW), IEEE. pp. 531–537.

[9] Khalil Zadeh, N., Sepehri, M.M., Farvaresh, H., 2014. Intelligent sales prediction for pharmaceutical distribution companies: A data mining based approach. Mathematical Problems in Engineering 2014.

[10] Kravets, A., Al-Gunaid, M., Loshmanov, V., Rasulov, S., Lempert, L., 2018. Model of medicines sales forecasting taking into account factors of influence, in: Journal of Physics:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

Conference Series, IOP Publishing. p. 032073.

[11] Krishna, A., Akhilesh, V., Aich, A., Hegde, C., 2018. Sales-forecasting of retail stores using machinelearning techniques, in: 2018 3rd International

[12] Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), IEEE. pp. 160–166.

[13] Lingxian, Y., Jiaqing, K., Shihuai, W., 2019. Online retail sales prediction with integrated framework of k-mean and neural network, in: Proceedings of the 2019 10th International Conference on E- business, Management and Economics, pp. 115–118.

[14] Loureiro, A.L., Miguéis, V.L., da Silva, L.F., 2018. Exploring the use of deep neural networks for salesforecasting in fashion retail. Decision Support Systems 114, 81–93.

[15] Onyema, E.M., Dalal, S., Romero, C.A.T. *et al.* Design of Intrusion Detection System based on Cyborg intelligence for security of Cloud Network Traffic of Smart Cities. Springer *J. Cloud Comp* **11,** 26 (2022). https://doi.org/10.1186/s13677-022-00305-6

[16] Pao, J.J., Sullivan, D.S., 2017. Time series sales forecasting. Final year project, Computer Science,Stanford Univ., Stanford, CA, USA .

[17] Park, J., Chang, B., Mok, N., 2019. 144 time series analysis and forecasting daily emergency department visits utilizing facebook's prophet method.Annals of Emergency Medicine 74, S57.

[18] Pavlyshenko, B.M., 2019. Machine-learning models for sales time series forecasting. Data 4, 15.

[19] Permatasari, C.I., Sutopo, W., Hisjam, M., 2018. Sales forecasting newspaper with arima: A casestudy, in: AIP Conference Proceedings, AIP Publishing LLC. p. 030017.

[20] Ribeiro, A., Seruca, I., Durão, N., 2017. Improving organizational decision support: Detection of outliers and sales prediction for a pharmaceutical distribution company. Procedia computer science121, 282–290.

[21] Sapankevych, N.I., Sankar, R., 2009. Time series prediction using support vector machines: a survey.IEEE computational intelligence magazine4, 24–38.

[22] Scherer, M., 2018. Multi-layer neural networks for sales forecasting. Journal of Applied Mathematics and Computational Mechanics 17, 61–68. Taylor, S.J., Letham, B., 2018. Forecasting at scale. The American Statistician 72, 37–45.

[23] Tyralis, H., Papacharalampous, G.A., 2018. Large-scale assessment of prophet for multi-step aheadforecasting of monthly streamflow. Advances in Geosciences 45, 147–153.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022**

[24] VAN Ruitenbeek, R., 2019. Hierarchical agglomerative clustering for product sales forecasting.

[25] Wang, H., 2020. An insurance sales prediction model based on deep learning. Rev. d'Intelligence Artif.34, 315–321.

[26] Zhang, W., Xu, Y., 2019. Fitting and prediction of total retail sales of consumer goods based on consumption indicators, in: The 4th International Conference on Economy, Judicature, Administrationand Humanitarian Projects (JAHP 2019), Atlantis Press. pp. 381–388.

[27] Onyema EM, Khalid K. Almuzaini, Onu FU, Devvret V, Ugboaja SG, Monika P, Afriyie RK, "Prospects and Challenges of Using Machine Learning for Academic Forecasting", Computational Intelligence and Neuroscience, vol. 2022, Article ID 5624475, 7 pages, 2022. https://doi.org/10.1155/2022/5624475

[28] Liu, Z., Wang, Y., & Feng, J. (2022). Vehicle-type strategies for manufacturer's carsharing. Kybernetes, ahead-of-print(ahead-of-print). doi: 10.1108/K-11-2021-1095

[29] Jia, T., Cai, C., Li, X., Luo, X., Zhang, Y.,... Yu, X. (2022). Dynamical community detection and spatiotemporal analysis in multilayer spatial interaction networks using trajectory data. International Journal of Geographical Information Science, 1-22. doi: 10.1080/13658816.2022.2055037

[30] Yang, D., Zhu, T., Wang, S., Wang, S., & Xiong, Z. LFRSNet: A Robust Light Field Semantic Segmentation Network Combining Contextual and Geometric Features. Frontiers in Environmental Science, 1443. doi: 10.3389/fenvs.2022.996513.

[31] Cao, B., Zhang, J., Liu, X., Sun, Z., Cao, W., Nowak, R. M.,... Lv, Z. (2021). Edge-Cloud Resource Scheduling in Space-Air-Ground Integrated Networks for Internet of Vehicles. IEEE internet of things journal, 1. doi: 10.1109/JIOT.2021.3065583

[32] Zhang, Z., Luo, C., & Zhao, Z. (2020). Application of probabilistic method in maximum tsunami height prediction considering stochastic seabed topography. Natural hazards (Dordrecht). doi: 10.1007/s11069-020-04283-3

[33] Zhang, Y., Liu, F., Fang, Z., Yuan, B., Zhang, G.,... Lu, J. (2021). Learning From a Complementary-Label Source Domain: Theory and Algorithms. IEEE transaction on neural networks and learning systems, PP, 1-15. doi: 10.1109/TNNLS.2021.3086093

[34] Yang, L., Xiong, Z., Liu, G., Hu, Y., Zhang, X.,... Qiu, M. (2021). An Analytical Model of Page Dissemination for Efficient Big Data Transmission of C-ITS. IEEE transactions on intelligent transportation systems, 1-10. doi:10.1109/TITS.2021.3134557

[35] Zenggang, X., Xiang, L., Xueming, Z., Sanyuan, Z., Fang, X., Xiaochao, Z.,... Mingyang, Z. (2022). A Service Pricing-based Two-Stage Incentive Algorithm for Socially Aware Networks. Journal of Signal Processing Systems. doi: 10.1007/s11265-022-01768-1

[36] Wu, X., Zheng, W., Xia, X., & Lo, D. (2021). Data Quality Matters: A Case Study on Data Label Correctness for Security Bug Report Prediction. IEEE transactions on software engineering, 1. doi: 10.1109/TSE.2021.3063727

[37] Zheng, W., Xun, Y., Wu, X., Deng, Z., Chen, X.,... Sui, Y. (2021). A Comparative Studyof Class Rebalancing Methods for Security Bug Report Classification. IEEE transactions on reliability, 70(4), 1-13. doi: 10.1109/TR.2021.311802