Fast and Accurate Supervised Machine Learning Strategy for Sales Prediction Using Real Time Datasets

1G.Sudha Gowd, Asst.Prof, Dept of CSE, Srinivasa Ramanujan Institute of Technology (A), Anantapur
2C.Nagesh, Asst.Prof, Dept of CSE, Srinivasa Ramanujan Institute of Technology (A), Anantapur
3 Tammineni Taruni, Asst.Prof, Dept of CSE, JNTUA, Anantapur
4 N.Kiran Kumar, Asst.Prof, Dept of CSE, JNTUA, Anantapur
5 G Pradeep Reddy, Asst.Prof, Dept of CSE, JNTUA, Anantapur

Abstract – *Restaurants require precise sales forecasting in order to encourage proper employee scheduling for crew load management. Using actual sales data from a mid-sized restaurant, this paper proposes a case study of numerous machine learning (ML) models. In vogue repetitive brain organization (RNN) and SVM models are incorporated for direct correlation with numerous techniques. To test the impacts of pattern and irregularity, we create three distinct datasets to prepare our models with and to analyze our outcomes. We engineer a lot of features and show how to choose the best subset of highly correlated features to help with forecasting. We look at the models in light of their presentation for guaging time steps of one-day and one-week over an organized test dataset. When it comes to one-day forecasting, linear models with a sMAPE of just 19.8% produce the best results. With errors below 20%, both ensemble models and two RNN models, LSTM and SVM, performed well. Non-RNN models performed poorly when forecasting for one week, delivering results with an error of more than 25%.*

Keywords: RNN, LSTM, SVM

1. Introduction

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).
In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc. In supervised learning, models are trained using labeled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

Supervised Learning:

First Determine the type of training dataset Collect/Gather the labeled training data. Split the training dataset into training dataset, test dataset, and validation dataset. Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output. Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.

Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets. Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

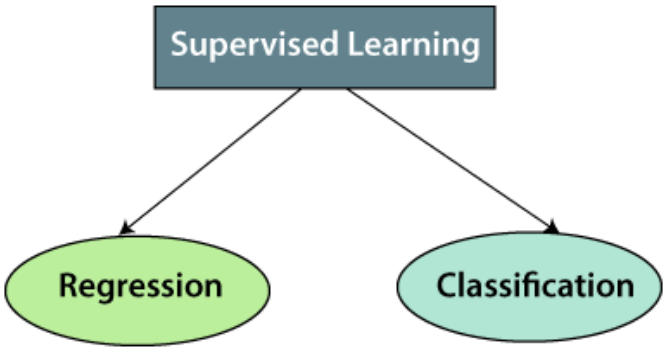

Fig.1.Supervised Learning

1. **Regression**

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

2. **Classification**

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Spam Filtering,

- Random Forest
- Decision Trees
- Logistic Regression
- Support vector Machines

In the modern landscape, neural networks and other machine learning methods have been suggested as powerful alternatives to traditional statistical analysis [5,6,7,8,9]. There are hundreds [10] of new methods and models being surveyed and tested, many of which are deep learning neural networks, and progress is being seen in image classification, language processing, and reinforcement learning [5]. Even convolutional neural networks have been shown to provide better results than some of the ARIMA models [6]. Traditionally, critics have stated that many of these studies are not forecasting long enough into the future, nor do they compare enough to old statistical models instead of trendy machine learning algorithms. Following, machine learning techniques can take a long time to train and tend to be 'black boxes' of information [10].

Although some skepticism has been seen towards neural network methods, recurrent networks are showing improvements over ARIMA and other notable statistical methods.

Especially when considering the now popular recurrent LSTM model, we see improvements when comparing to ARIMA models [8,9], although the works do not compare the results with a larger subset of machine learning methods. Researchers have recently begun improving the accuracy of deep learning forecasts over larger multi-horizon windows and are also beginning to incorporate hybrid deep learning-ARMIA models [7]. Safe lengths of forecast horizons and techniques for increasing the forecasting window for recurrent networks are of particular interest [11]. Likewise, methods for injecting static features as long-term context have resulted in new architectures which implement transformer layers for short-term dependencies and special self-attention layers to capture long-range dependencies [5].

2. Research Survey

The idea is that the restaurant will be able to schedule employees at a lower cost if it can accurately predict sales. Typically, the person making the schedule intuitively completes this forecasting task, and sales averages frequently aid in the prediction. To schedule employees, managers do not need to be aware of the minute-to-minute sales figures. In this way, we center around finding segments of times representatives are working, for example, dayshift, center shift, and nightshift. Predictions need to be made one week in advance in order to be useful in the real world because no restaurant schedules employees one day at a time. Guest counts and sales dollars, which should be forecasted with high accuracy, have been identified as the most important forecasted criteria, according to empirical evidence gathered through interviews with retail managers [1]. These kinds of predictions are typically made in one of three ways in restaurants: 1) through the sound judgment of a manager; 2) through economic modeling; or 3) through time series analysis [2].

A comparative café writing survey on a few models/eateries [3] shows how the information is arranged will profoundly impact the technique utilized. Numerous statistical, machine learning, and deep learning models yield positive results; however, the "No Free Lunch" theorem predicts that each model will have some drawbacks [3]. A subjective report was directed in 2008 on seven deeply grounded café networks in a similar region as the eatery for our situation study. The chains sold between $75 million and $2 billion and had between 23 and 654 restaurants. The majority of forecasting methods utilized statistical or regression techniques, but neither ARIMA nor neural networks were utilized [4]. The fact that ARIMA models are no longer used to model complex time series problems provides a solid foundation for this study to determine whether advancements in neural network research have made them useful for restaurant forecasting.

This knowledge Discovery in Databases (KDD) procedure consists of a series of the following steps [1] Records cleansing – To do away with noise and beside the point information. [2] Information integration – In which a couple of facts assets are combined. [3] Information choice – For retrieving from the database most effective the applicable facts for the evaluation. [4] Facts transformation – Where in statistics are converted or consolidated into appropriate bureaucracy for mining. [5] Statistics mining – The section wherein the algorithms are implemented so as to extract information styles. [6] Sample evaluation – To discover the thrilling patterns which represents new knowhow. [7] Know-how presentation – When the visualization strategies are used to offer the mined information to the consumer.

The comparison of ML techniques to forecast curated restaurant sales is a common research question and can be seen in several recent works [14,15,16,17]. Two additional recent, non-restaurant forecasting with ML problems are also examined [11,18]. Although similar in model training and feature engineering techniques, our methodology differs from other recent

forecasting papers in a few key areas, which we outline. The first important difference in researched methods is the forecasting horizon window used. Many papers either used an unclear horizon window or made forecasts of only one time step at a time [14,15,16,17,18].

Only one paper increased the forecast horizon beyond one time step [11], so we consider forecasting one week of results the main contribution of this research. Another point of departure with reviewed papers is the importance of stationary data. Traditionally, it is important to have a stationary dataset when working with time series so there is no trend or seasonality learned—instead, each instance is separated and can be forecasted based on its own merit instead of implicit similarity. However, only one paper [11] even mentions this stationary condition. Instead of exploring it further, the authors simply trained models using data that did not seem to have any associated trend. As an extension to these works, we consider the stationary condition and test multiple datasets to gauge its importance.

Ellero and Pelegrini (Ellero and Pellegrini, 2014) [2] assess the performance of various widely-adopted models from literature to forecast Italian occupancy rate . They find that exponential smoothing, advanced pick-up, and moving average models show the simplest success within the compared models. Shenoy et al.(Shenoy et al., 2017)[3] demonstrate their estimation of reservation information supported user activity and search results using the info provided by Expedia. Their studies show that acquisition of serious results becomes possible through clustering and ensemble operations. Song and Li (2007)[4] included in their review was “A practitioners guide to time-series methods for tourism demand forecasting -a case study of Durnban, South Africa ” by Burger, Dohnal, Kathrada and Law (2001). the target of the study was to conduct a forecast folks demand for visit Durban, South Africa. during a review of the planet Tourism Organization in 1995 about African tourism, South Africa was considered to be "one of the foremost promising tourism destinations on the African continent" but it's not been ready to realize its full potential yet.

3. Approaches for Implementation

Methods of partitioning, feature selection, and differencing are studied uniquely for this data but also are listed as general techniques to follow for other similar problems or datasets. We briefly list all the methods used for the forecasting task. Each model goes through a specific training pipeline which will be described in some detail. Next, we describe the recurrent neural network architectures used in this work in enough detail to provide context for those unfamiliar. Following, we describe the metrics used in this study to compare the various modeling techniques employed. Finally, we discuss how models are compared against common baseline approaches, filtered to acquire the best feature subset, then tested to acquire our final results.

Classification The way toward ordering data sets into outright gatherings with one another. The delegates of each club are "as close as could reasonably be expected" to one another, the various clubs are "far" from one another, and separation is, for instance, the specific variable you are attempting to foresee. For instance, a common arrangement multifaceted nature is to separate an organization's database into clubs that are as uniform as conceivable concerning the acknowledge quality factors for values of "Great" and "Awful".

The proposed framework expect that the organization's three branches and its regarded branch administrator can get to the branch database and find shrouded designs in the database. Rather than having three databases at each branch that store all organization records, there is a focal server, kept up by an overseer. A fascinating inquiry is the thing that entrance to make. The two directors and heads can change the data. That is, you can include and erase records in the database. To discover concealed examples in the database, they utilize a framework that

actualizes a from the earlier algorithm to discover visit things in the database and can see the yield as reports and pie diagrams. Pie outlines are utilized for a superior comprehension. In this way, the supervisor of each branch can locate the successive things in that branch, since the overseer needs to settle on significant choices about the organization, and the manager can see or search the regular things in all the branches.

In the year 2019, authors Lim et al. released the temporal fusion transformer (TFT) network, a novel, attention-based learning architecture that combines high-performance multi-horizon forecasting with interpretable insights into temporal dynamics [5]. We give a brief description of the TFT model as an introduction as the model architecture is laid out in sufficient detail in the preliminary paper titled Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting.

Recurrent neural networks are one of the many variations of ANN's. RNN's are common amongst tasks such as speech recognition, translation but also timeseries forecasting [12]. The difference between RNN's and regular feedforward neural networks (FFNN) are that RNN's data moves in both directions [13]. Therefore, decisions made in an RNN are based on both the current input and previous predictions, with this in consideration it makes clear how the application of RNN to sequential data is relevant. In deep neural networks Jürgen Schmidhuber introduces Credit Assignment Paths (CAP) [14], which is a system for analysing and classifying the depth of a neural network related to the depth of a problem. In this regard RNN's, due to the reuse of previous data can potentially solve problems of unlimited depth whereas in feed forward neural networks increasing the depth of the model in relation to the depth of the problem is more relevant.

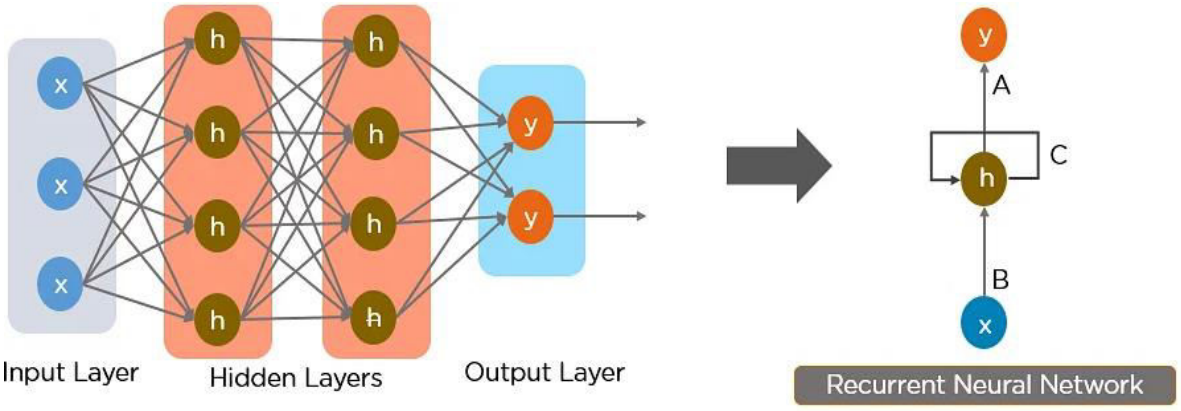

Fig.2. Recurrent neural network

Long short-term memory is a method introduced by Schmidhuber and Hochreiter in 1997 [15]. LSTM is an RNN architecture that consists of additional gates that help the neural network adapt to error values from the output layer when backpropagating. They found many advantages of LSTM in comparison to previous solutions. They came to the conclusion that LSTM was efficient in a broad range of problems and completely solved the issue with backpropagation through time (BPTT). Briefly backpropagation is a method used to calculate the weights of an artificial network to increase performance of training, in an RNN with LSTM BPTT is used. BPTT uses stochastic gradient descent to calculate the weights and therefore the relevance of a non-vanishing gradient becomes essential. For further reading see [16].

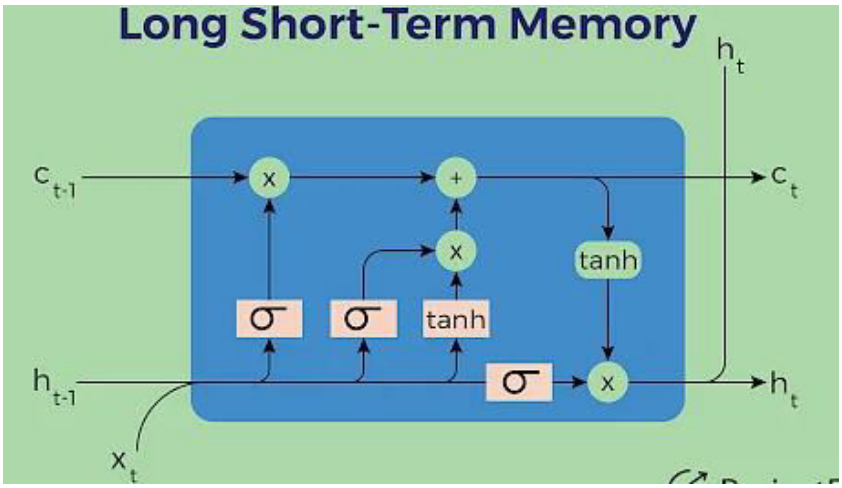

Fig.3. Long short-term memory Architecture

The TFT architecture uses a direct forecasting method [7] where any prediction has access to all available inputs and achieves this uniquely by not assuming that all time-varying variables are appropriate to use [5]. The main components of the architecture are gating mechanisms, variable section networks, static covariant encoders, a multi-head attention layer, and the temporal fusion decoder. The authors propose a gating mechanism called Gated Residual Network (GRN), which may skip over unneeded or unused parts of the model architecture. Variable selection network ensures relevant input variables are captured for each individual time step by transforming each input variable into a vector matching dimensions with subsequent layers. Each static, past, and future inputs acquire their own network for instance-wise variable selection.

Static variables, such as the date or a holiday, are integrated into the network through static enrichment layers to train for temporal dynamics properly. The static covariant encoder integrates information from static metadata to be used to inform context for variable selection, processing of temporal features, and enrichment of those features with the static information [5]. Short-term dependencies are found with LSTM layers, and long-term dependencies are captured with multi-headed self-attention block layers. An additional layer of processing is completed on the self-attention output in the position-wise feed-forward layer. This process is designed such as the static enhancement layer.

Support Vector Machine (SVM) is a supervised machine learning method developed by Cortes and Vapnik in 1995 used for binary classification [19]. It has since then been developed for use in regression and outlier detection [20]. SVM classifies data in different groups with support vectors used as divisors in a multidimensional space. Support vectors are hyperplanes which is a subspace of one dimension lower than its ambient space [21]. Support vectors are calculated with a kernel function, the three most popular kernel functions are Linear, Polynomial and Radial Basis Function (RBF), however there are several other kernel functions. To classify a linear classification problem a linear kernel function is recommended [21]. Linear kernel functions calculate linear support vectors that are placed in the multidimensional space with a maximum calculated space from different classification group elements. Nonlinear classification problems require a nonlinear kernel function such as polynomial- or RBF- kernel function. Polynomial kernels work in a similar way as linear functions, except they use a polynomial

function that produces polynomial hyperplane curves as divisors in the multidimensional space. RBF kernels nonlinearly maps samples into a multidimensional space [22]. This kernel is usually time consuming due to the fact that it maps every sample in the multidimensional space.

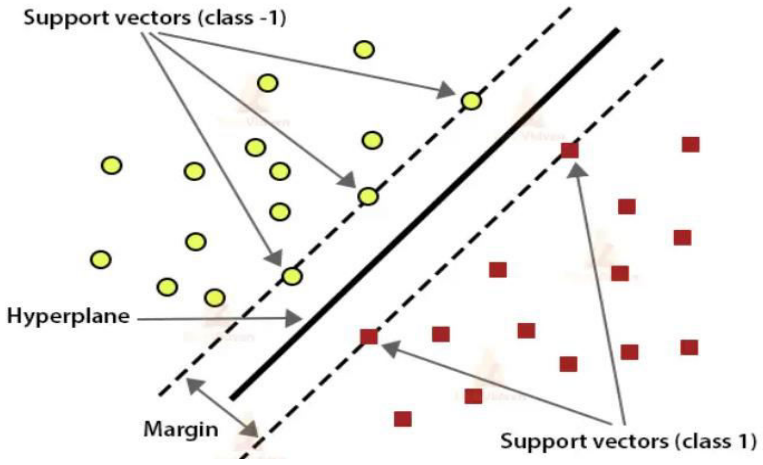

Fig.4. Support Vector Machine

4. Results

From the Tables I, II , it has been observed that the Mean Absolute Error is highest in case of Support Vector Regression for all the three years and minimum in case of Extra Tree Regression. Mean Absolute Error is the average magnitude of the error in prediction set. It is the average over the test sample of absolute difference between prediction and actual observation.

TABLE I. STATISTICAL MEASURES FOR DATASET OF YEAR 2010

Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	424421.93	251400879887.53	501398.92
Random Forest Regression	80396.14	25126954749.18	158514.84
KNN Regression	281135.23	131980791838.85	363291.60
Support Vector Regressor	470857.85	320200129812.73	565862.28
Extra Tree Regression	48281.35	5534368506.39	74393.33

TABLE II. STATISTICAL MEASURES FOR DATASET OF YEAR 2011

Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	409909.51	235782880216.47	485574.79
Random Forest Regression	43811.57	4031635222.35	63495.15
KNN Regression	272092.75	123596844025.18	351563.42
Support Vector Regressor	430737.43	267243666430.57	516956.15
Extra Tree Regression	42752.07	3840250218.75	61969.75

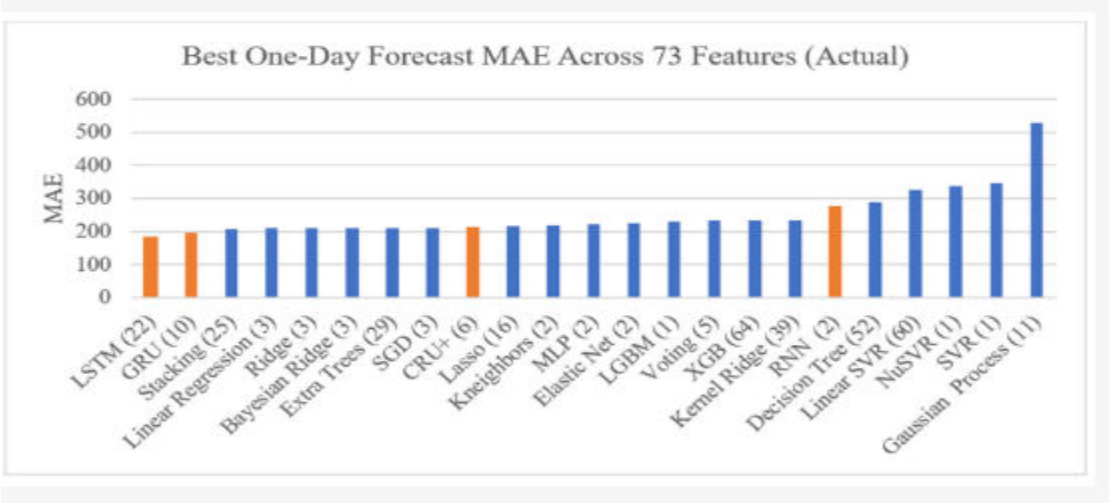

Fig.5.Forecasting of MAE

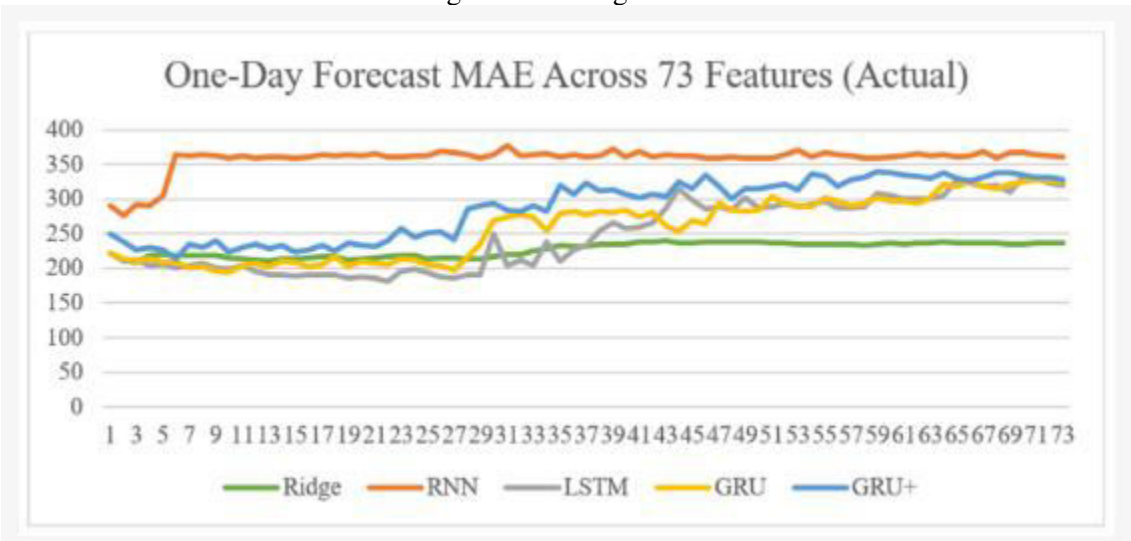

Fig.6.Forecasting of MAE as aggregate

5. Conclusion

Three datasets were tested using over 20 models to show the impact of creating stationary datasets on the feature engineering and model training processes. In the end, recurrent neural network (RNN) models performed better than other models, with the SVM model showing interesting results due to static context injection. SVM only performed slightly worse in one-day forecasting when compared to the best performing ridge regression model using the daily differenced dataset, but with current methods, the ridge model could not scale to one-week forecasting. While other models performed well with the differenced datasets, the RNN models struggled to provide good results in either one-day and one-week testing, so it is not recommended to use differencing techniques. Even when learning on the previous 14 days of targets, a single instance just does not give enough context to allow great forecasts that far into the future in one step. Transformer encoding and decoding learning layers show potential for improvements in forecasting problems. . With our research, we have shown that the enhancements added to the SVM model over basal RNN models allow multi-horizon predictions

well into the future. Other models made comparable results in one-day and one-week forecasting, but no other model can as reliably forecast zero-sale days, such as holidays, from only a few samples. Even though the SVM model was outperformed in single one-day forecasting, the ability to give or withhold context at key moments provides for more robust predictions which scale better into the future.

References:

[1] TianLinqin, Application of Data Mining Technology in Tobacco Industry, Journal of Agricultural Science and Technology, March ,2012.

[2] P.Naresh,et.al., "Implementation of Map Reduce Based Clustering for Large Database in Cloud", Journal For Innovative Development in Pharmaceutical and Technical Science,vol.1,pp 1-4,2018.

[3] Wang Ying, Li Renwang, Li Bin, Zhang Zhile, Costume Sales Forecasting Model Based on CURE Algorithm and C4.5 Decision Tree, Journal of Textile Research, September, 2008.

[4] Zhang Gefu, Ou yang, Hao nan, Xu qi, Application of Decision Tree in Apparel Marketing Based on Appearance of Consumers, Journal of Computer Applications, July,2010.

[5] Schuster and R. Wolff, "Communication-Efficient Distributed Mining of Association Rules", Proc. ACM SIGMOD Int'l Conf. Management of Data, ACM Press, 2001,pp. 473-484.

[6] Kimball R., Ross M.:, "The Data Warehouse Toolkit, The Complete Guide to Dimensional Modeling", 2nd edn. John Wiley & Sons, New York (2002).

[7] P, N., & R Suguna. (2022). Enhancing the Performance of Association Rule Generation over Dynamic Data using Incremental Tree Structures. International Journal of Next-Generation Computing, 13(3). https://doi.org/10.47164/ijngc.v13i3.806.

[8] M.Z Ashrafi, Monash University ODAM:, "An Optimized Distributed Association Rule Mining Algorithm", IEEE DISTRIBUTED SYSTEMS ONLINE 1541-4922 © 2004.

[9] Ma, Y., Liu, B., Wong, C.K.: Web for Data Mining:, "Organizing and Interpreting the Discovered Rules Using the Web", SIGKDD Explorations, Vol. 2 (1). ACM Press, (2000) 16-23.

[10] R. Agrawal and J.C. Shafer , "Parallel Mining of Association Rules", IEEE Tran. Knowledge and 16Data Eng. , vol. 8, no. 6, 1996,pp. 962- 969;

[11] Naresh, P., & Suguna, R. (2019). Association Rule Mining Algorithms on Large and Small Datasets: A Comparative Study. 2019 International Conference on Intelligent Computing and Control Systems (ICCS). DOI:10.1109/iccs45141.2019.9065836.

[12] P. Naresh, K. Pavan kumar, and D. K. Shareef, 'Implementation of Secure Ranked Keyword Search by Using RSSE,' International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 3, March – 2013.

[13] M. I. Thariq Hussan, D. Saidulu, P. T. Anitha, A. Manikandan and P. Naresh (2022), Object Detection and Recognition in Real Time Using Deep Learning for Visually Impaired People. IJEER 10(2), 80-86. DOI: 10.37391/IJEER.100205.

[14] Suguna Ramadass and Shyamala Devi 2019 Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers Springer's book series Learning and Analytics in Intelligent Systems, Springer.

[15] Naresh P, Shekhar GN, Kumar MK, Rajyalakshmi P. Implementation of multi-node clusters in column oriented database using HDFS. Empirical Research Press Ltd. 2017; p. 186.

[16] V.Krishna, Dr.V.P.C.Rao, P.Naresh, P.Rajyalakshmi " Incorporation of DCT and MSVQ to Enhance Image Compression Ratio of an image" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 03 Issue: 03 | Mar-2016.

[17] B.M.G. Prasad, P. Naresh, V. Veeresh, "Frequent Temporal Patterns Mining With Relative Intervals", International Refereed Journal of Engineering and Science ,Volume 4, Issue 6 (June 2015), PP.153-156.

[18] T. Aruna, P. Naresh, A. Rajeshwari, M. I. T. Hussan and K. G. Guptha, "Visualization and Prediction of Rainfall Using Deep Learning and Machine Learning Techniques," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 910-914, doi: 10.1109/ICTACS56270.2022.9988553.

[19] V. Krishna, Y. D. Solomon Raju, C. V. Raghavendran, P. Naresh and A. Rajesh, "Identification of Nutritional Deficiencies in Crops Using Machine Learning and Image Processing Techniques," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 925-929, doi: 10.1109/ICIEM54221.2022.9853072.

[20] B. Narsimha, Ch V Raghavendran, Pannangi Rajyalakshmi, G Kasi Reddy, M. Bhargavi and P. Naresh (2022), Cyber Defense in the Age of Artificial Intelligence and Machine Learning for Financial Fraud Detection Application. IJEER 10(2), 87-92. DOI: 10.37391/IJEER.100206.

[21] Naresh, P., & Suguna, R. (2021). IPOC: An efficient approach for dynamic association rule generation using incremental data with updating supports. Indonesian Journal of Electrical Engineering and Computer Science, 24(2), 1084. https://doi.org/10.11591/ijeecs.v24.i2.pp1084-1090.

[22] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," Proc. ACMSIGMOD Int'l Conf. Management of Data, , May 1993.