BIG-MART SALES PREDICTION ANALYSIS

Team Details

Pratheek Babu	PES1UG20CS674
Pruthvi S Kodmurgi	PES1UG20CS676
Vijay Ram Enaganti	PES1UG20CS700
Richa Ramesh	PES1UG20CS712

Literature Survey

Paper - 1: Time-series forecasting of seasonal items sales using machine learning – A comparative analysis

Yasaman Ensafi , Saman Hassanzadeh Amin , Guoqing Zhang , Bharat Shah

Introduction:

To this aim, several forecasting models are applied. First, some classical time-series forecasting techniques such as Seasonal Autoregressive Integrated Moving Average (SARIMA) and Triple Exponential Smoothing are utilized. Then, more advanced methods such as Prophet, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) are applied. The performances of the models are compared using different accuracy measurement methods (e.g., Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE)). The results show the superiority of the Stacked LSTM method over the other methods.

Dataset:

Dataset contains 9994 data points, 21 features, and zero null values. The columns are attributes such as Order ID, Ship Date, Order Date, Ship Mode, Customer Name, Customer ID, Segment, City, Country, State, Region, Postal Code, Category, Product Name, Sub-Category, Sales, Discount, Quantity, and Profit.

Aim:

The aim of this study is to perform a time-series forecasting on a seasonal item with the help of different models such as SARIMA, Exponential Smoothing, Prophet, and Neural Networks

Methodology:

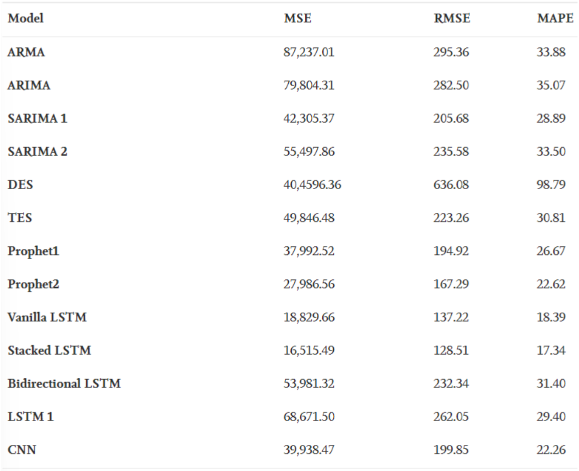
Forecasting Models:

- ARMA, ARIMA and SARIMA
- Single, Double and Triple Exponential Smoothening
- Prophet Model by Facebook (Meta)
- ANN
    - RNN
    - LSTM
        - LSTM
        - Vanilla LSTM
        - Stacked LSTM
        - Bidirectional LSTM
    - CNN

Metrics:

- MSE
- RMSE
- MAPE

Results:



Model	MSE	RMSE	MAPE
ARMA	87,237.01	295.36	33.88
ARIMA	79,804.31	282.50	35.07
SARIMA 1	42,305.37	205.68	28.89
SARIMA 2	55,497.86	235.58	33.50
DES	40,4596.36	636.08	98.79
TES	49,846.48	223.26	30.81
Prophet1	37,992.52	194.92	26.67
Prophet2	27,986.56	167.29	22.62
Vanilla LSTM	18,829.66	137.22	18.39
Stacked LSTM	16,515.49	128.51	17.34
Bidirectional LSTM	53,981.32	232.34	31.40
LSTM 1	68,671.50	262.05	29.40
CNN	39,938.47	199.85	22.26

After performing sales forecasting with various methods, the following conclusions can be drawn from this study. First, most of the neural network methods have performed better than the classical forecasting methods. In addition, the results have indicated the superiority of Stacked LSTM over the other models. Moreover, the CNN method which is usually applied for the task of image recognition showed a great performance in this case.

Paper - 2: Machine Learning Outperforms Classical Forecasting on Horticultural Sales Predictions

Florian Haselbeck , Jennifer Killinger , Klaus Menrad , Thomas Hannus , Dominik G. Grimm

Introduction:

Forecasting on horticultural sales predictions is essential for businesses in the horticulture industry. Accurate sales forecasts can help businesses plan their production and supply chain, optimize inventory levels, and make informed decisions about pricing and marketing strategies. They can also help businesses manage risks and mitigate the effects of external factors such as weather and market trends.

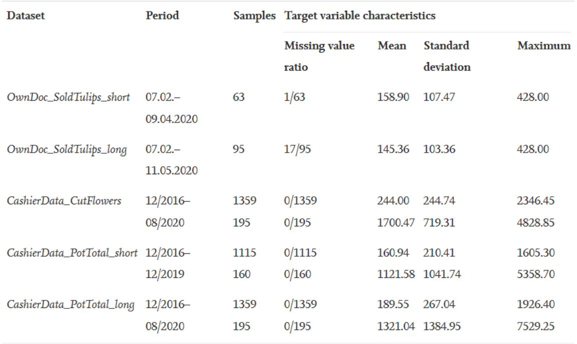
Furthermore, sales forecasts can provide valuable insights into consumer demand and market trends, allowing businesses to identify opportunities and adapt to changes in the market. By providing a forecast of future sales, businesses can make more informed decisions about investment, expansion, and other strategic initiatives.

In summary, forecasting on horticultural sales predictions is crucial for the success and growth of businesses in the horticulture industry. It enables businesses to make informed decisions, manage risks, and capitalize on opportunities in the market.

For this purpose, the performance of 9 state-of-the-art machine learning models were compared with the performance of 3 classic forecasting models. Results show that the machine learning models are superior to classic models with XGBoost being the best among them.

Dataset:

typical horticultural retail sales data from Germany.



Dataset	Period	Samples	Target variable characteristics			
			Missing value ratio	Mean	Standard deviation	Maximum
OwnDoc_SoldTulips_short	07.02.–09.04.2020	63	1/63	158.90	107.47	428.00
OwnDoc_SoldTulips_long	07.02.–11.05.2020	95	17/95	145.36	103.36	428.00
CashierData_CutFlowers	12/2016–08/2020	1359 195	0/1359 0/195	244.00 1700.47	244.74 719.31	2346.45 4828.85
CashierData_PotTotal_short	12/2016–12/2019	1115 160	0/1115 0/160	160.94 1121.58	210.41 1041.74	1605.30 5358.70
CashierData_PotTotal_long	12/2016–08/2020	1359 195	0/1359 0/195	189.55 1321.04	267.04 1384.95	1926.40 7529.25

Methodology:

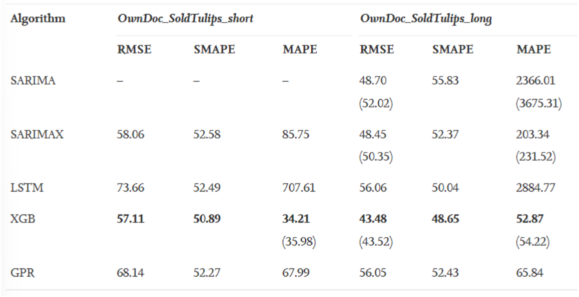
- Preprocessing

  Many algorithms cannot handle missing values, except for XGB. Therefore, we applied different strategies for data imputation: Mean, K-Nearest-Neighbors (KNN) and Iterative Imputation.

- Model Selection
  - ES and SARIMA - Univariate Techniques
  - SARIMAX - Multivariate Technique
  - MLR - Baseline Regression
  - ML Models
    - ANN
    - RNN with LSTM
    - XGBoost
    - GPR
- Evaluation Metrics
  - RMSE
  - sMAPE
  - MAPE

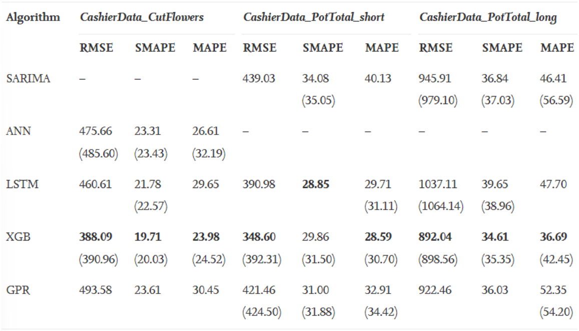
Results and Conclusion:

The top 5 best performing algorithms for OwnDoc dataset



Algorithm	*OwnDoc_SoldTulips_short*			*OwnDoc_SoldTulips_long*		
	RMSE	SMAPE	MAPE	RMSE	SMAPE	MAPE
SARIMA	–	–	–	48.70 (52.02)	55.83	2366.01 (3675.31)
SARIMAX	58.06	52.58	85.75	48.45 (50.35)	52.37	203.34 (231.52)
LSTM	73.66	52.49	707.61	56.06	50.04	2884.77
XGB	**57.11**	**50.89**	**34.21** (35.98)	**43.48** (43.52)	**48.65**	**52.87** (54.22)
GPR	68.14	52.27	67.99	56.05	52.43	65.84

The top 5 best performing algorithms for CashierData dataset



Algorithm	*CashierData_CutFlowers*			*CashierData_PotTotal_short*			*CashierData_PotTotal_long*		
	RMSE	SMAPE	MAPE	RMSE	SMAPE	MAPE	RMSE	SMAPE	MAPE
SARIMA	–	–	–	439.03	34.08 (35.05)	40.13	945.91 (979.10)	36.84 (37.03)	46.41 (56.59)
ANN	475.66 (485.60)	23.31 (23.43)	26.61 (32.19)	–	–	–	–	–	–
LSTM	460.61	21.78 (22.57)	29.65	390.98	**28.85**	29.71 (31.11)	1037.11 (1064.14)	39.65 (38.96)	47.70
XGB	**388.09** (390.96)	**19.71** (20.03)	**23.98** (24.52)	**348.60** (392.31)	29.86 (31.50)	**28.59** (30.70)	**892.04** (898.56)	**34.61** (35.35)	**36.69** (42.45)
GPR	493.58	23.61	30.45	421.46 (424.50)	31.00 (31.88)	32.91 (34.42)	922.46	36.03	52.35 (54.20)

Machine learning models have several advantages over classic forecasting models. One advantage is that they can automatically learn and improve from data, without the need for explicit programming. This allows them to be more flexible and adapt to a wider range of situations. Additionally, machine learning models can often handle large and complex datasets more effectively than classic models, which can make them more accurate and reliable. Finally, machine learning models are able to identify nonlinear relationships in data that classic models may not be able to detect, which can improve the accuracy of their predictions.

Paper 3: Exploratory Data analysis and sales forecasting of bigmart dataset using supervised and ANN algorithms (Base Paper)

T.K. Thivakaran , M. Ramesh

Introduction:

Sales forecasting is the process of predicting future sales for a given product or service. This is important for businesses as it helps them make informed decisions about production, inventory, and pricing. In this report, we will explore the use of supervised and ANN algorithms for sales forecasting using the Bigmart dataset.

The Bigmart dataset contains various features and variables related to the sales of different products across different outlets. This includes information such as product type, price, and location. We will use this dataset to train and test machine learning models that can accurately predict sales for the future.

First, we will use a supervised learning algorithm, such as linear regression or decision trees, to train the model on the dataset. This will allow the model to learn the relationships between the various features and the target variable (sales). We will then evaluate the performance of the model and make predictions on the testing data.

If the performance of the supervised learning model is not satisfactory, we will try using an ANN algorithm instead. ANNs are more complex and can handle larger and more complex data sets, potentially leading to more accurate predictions. We will train the ANN model on the dataset and compare its performance to the supervised learning model.

Finally, we will use the best-performing model to make predictions on the entire Bigmart dataset and generate a sales forecast for the future. This will provide valuable insights for businesses and help them make better decisions.

Dataset:

A dataset of wide market sales data, which has 12 features. These 12 criteria define the fundamental characteristics of the data being projected. Answer variables and predictors are two types of qualities. A dataset that contains 8523 objects from various places and cities.

Methodology:

- Preprocessing:

  Replace the missing value for outlet size with the mode value of that attribute, and we substitute the mean value for the missing values of that particular property of object weight.

- Models:

  1. Regression
     a. Linear Regression
     b. Random Forest Regression
     c. Lasso Regression
     d. XGBoost Regression
     e. Ridge Regression
  2. Artificial Neural Networks (ANN)

- Metrics and Parameters:

  For Regression Models:

  1. Linear Regressor
     a. Loss = Mean Absolute Error
  2. XGBoost Regressor
     a. N estimators = 1000
     b. Learning Rate = 0.05
     c. Loss = Mean Absolute Error
  3. Lasso Regressor
     a. Alpha = 0.01
     b. Loss = Mean Absolute Error
  4. Ridge Regressor
     a. Alpha = 1
     b. Loss = Mean Absolute Error

  For ANN Model:

  - Optimizer = Adam
  - Loss = Mean Absolute Error
  - Kernel Initializer = Normal
  - Epochs = 100
  - Batch Size = 64
  - Validation Split = 0.2
  - Activation Function
    - Relu and Linear

Result and Conclusion:

Exploratory machine learning approaches are used to anticipate bigmart store sales. The efficiency and accuracy of the proposed system will improve. The profit made by the company is exactly proportional to the prediction. The efficiency of the system can also be increased by investigating how the submodels function. Sales extrapolations are cheap and frequently sufficient for the necessary judgments.