Long Short-Term Memory-RNN based model for Multivariate Car Sales Forecasting

Preeti Saxena¹, Pritika Bahad²*, Raj Kamal³

¹² School of Computer Science and IT, Devi Ahilya University, Indore, India
³ Prestige Institute of Engineering, Management and Research, Indore, India
¹ preeti_ms@rediffmail.com, ² bahad.pritika@gmail.com, ³ dr_rajkamal@hotmail.com

Abstract

The paper presents a study of deep learning-based models for forecasting future directions of car sales, and car model preferences. An open-source Kaggle multivariate datasets for many years available for Norway new car sales. They are used for analyzing and predicting. The results based on Autoregressive Integration Moving Average (ARIMA) and Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN) based models are analyzed and used for forecasting future directions. The present study is useful for identifying features of different variants of all used (imported) cars, electric-used cars, and new diesel cars. The implementation results showed reduced Mean Absolute Error (MAE) and Root-Mean-Square Error (RMSE) for LSTM-RNN based time-series forecasting. The study forecasts the rise of green vehicles in the upcoming years in Norway. The performance of a model depends upon the characteristics of the dataset. The results show that LSTM-RNN is thus better than the ARIMA for the multivariate datasets. The interpretation of these results shows LSTM-RNN based time-series forecasting can be utilized for valuable forecasting in various domains such as credit, insurance, consumer behavior, and medical diagnosis.

Keywords: Data analytics, Deep Learning, Autoregressive Integration Moving-Average (ARIMA), Long Short-Term Memory Neural Network (LSTM), Time-Series Forecasting.

1. Introduction

Statistical forecasting applies statistics based methods on historical data to plan the future. This can be done on any quantitative data such as sales, stock, population, and GDP. Time-series analysis is a statistical technique that is commonly used for extrapolating past behavior into the future. Time-series analysis basically deals with time-series data or trend analysis. The time-series data provides visual information to the unpredictable future quantitative data.

The trained and tested models are used to predict new values or results [1] and classify [2] Autoregressive Integration Moving Average (ARIMA) is a statistical technique that uses time-series data to predict the future. ARIMA uses regression, a time-series method that uses basic statistics to predict future values for a target variable [3]. Artificial neural network (ANN) is a Deep Learning (DL) technique, which predict patterns in extremely complex problems. Time-series prediction with DL methods, especially Long Short-term Memory Neural Network (LSTM), has scored significant achievements in recent years. LSTM is a recurrent neural network (RNN) effective at * Corresponding Author
capturing long-term dependencies of complex time-series data. LSTM automatically determines the optimal time lags. There are many parameters such as the number of layers and nodes to choose, but also to improve the model and integration. After the training is complete, there are alternative models that must be evaluated on different aspects [4].

Efficient analytics requires a rich dataset. A rich dataset is needed to provide high-quality analytics results. A rich dataset means the one that offers a vast opportunity for exploration and offers an immense range of data patterns[1]. The datasets contain monthly car sales for 2007-2017 by the make and the data for most popular models[5]. Datasets from Norwegian Car-Sales Company open source from Kaggle, are rich and thus are chosen for analyzing and predicting future directions.

ARIMA and LSTM are used to forecast future directions. Also, some behavior of the car datasets enabled the interesting insights.

The remainder of this paper is organized as follows. Section 2 provides the background and motivations of this work. Section 3 gives an overview of forecasting techniques, ARIMA, and LSTM used in the experiments. Section 5 presents the dataset description, the experimentations, results, and their salient observations followed by conclusions in Section 5.

2. Background

Most importantly DL can be used for analysis and learning where raw data is largely unlabeled and un-categorized [6]. The DL models are also useful for discovering patterns and predicting the activities of the users [7].

Data analytics over large datasets requires open-source machine learning libraries and frameworks include Scikit-learn [8], Spark MLlib [9], and TensorFlow[10]. Scikit-learn is an open-source machine learning library that supports supervised and unsupervised learning. Spark is another open-source project of Apache foundation that is developed for big data processing. Spark MLlib is aimed at providing scalable ML facilities for Spark clusters. TensorFlow is an open-source framework developed by Google for carrying out high-performance numerical computations.

ARIMA model is used to predict on univariate financial time-series such as electricity prices in Sweden[11] and tomato prices in Serbia[12]. A non-linear approach for time-series forecasting is to use neural network-based models. Different neural network based models are used to predict on univariate and multivariate time-series data such as air quality prediction[13], house price predictions[14], and stock price predictions[15]. The comparative study suggests the use of different learning and optimization methods increase the accuracy of the predictions [16,17].

3. Forecasting Algorithms

The time-series techniques are used to forecast the data. The techniques include linear and non-linear algorithms. ARIMA models and their variations such as AR, MA, and ARMA fall under linear class. A non-linear approach for time-series forecasting is to use neural network based model. RNNs are able to capture sequential information by carrying results from previous computations or states into the next states. LSTM is a type of RNN with the capability of remembering the values from earlier stages for the purpose of future use. These networks can be trained on variable sized input and are
able to produce variable sized output, this makes them suitable for capturing temporal data and thus for the task of forecasting. Within the domain of forecasting, LSTM’s have been utilized to predict on univariate and multivariate financial time-series.

The current study compares ARIMA and LSTM models with respect to their performance in reducing error rates.

1) Time-series forecasting using ARIMA

The ARIMA model combines the AR process and the MA process, where the terms of the former are based directly on the previous values, and the terms of the latter on previous innovations [18]. Multiple regression models forecast a variable using a linear combination of predictors, while autoregressive models use a combination of past values of the variable. The parameters p and q are non-negative integers that regulate the number of terms, respectively. Finally, the model provides an integrated approach for transforming input time-series into stationary time-series, where the order of differencing is determined by the parameter d, this makes the model suitable for forecasting time-series in practice. The seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model. The model can be represented as

$$\text{ARIMA} \left(p,d,q\right) \times \left(P,D,Q\right)_S$$

where p= non-seasonal AR order, d= non-seasonal differencing, q= non-seasonal MA order, P= seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = time span of repeating seasonal pattern.

Steps to forecast using ARIMA

1. Check Stationarity: A time-series may have a trend or seasonality component. It must be made stationary to use ARIMA for forecast.
2. Difference: Differencing is performed to get rid of the varying mean. Iteratively perform differencing to stationarize the time-series.
3. Filter out a validation sample: The dataset is divided into training, testing and validation sets. Validation set is used to check accuracy of the model.
4. Select AR and MA terms: The autocorrelation function (ACF) and partial autocorrelation (PACF) are used to decide whether to include an AR term(s), MA term(s), or both.
5. Build the model: The model is build and depending on need the number of periods to forecast is set.
6. Validate model: The predicted values and the actual values in the validation set is compared.
7. Deploy the model for forecasting.

2) Time-Series Forecasting using Long Short-Term Memory-Recurrent Neural Network

RNN use back-propagation algorithm for training. During neural network training, the minimum of the error function is found by iteratively taking small steps in the direction of the negative error derivative with respect to networks weights. With each subsequent layer the magnitude of the gradients gets exponentially smaller. RNN suffers as a result
from vanishing gradient problem in the lower layers of a deep network.

Long Short-Term Memory network (LSTM) is a special kind of RNN competent of learning long-term dependencies[19]. LSTM is a very effective solution for addressing the vanishing gradient problem. The hidden layer of basic RNN is replaced by an LSTM cell in LSTM-RNN. Figure 1 shows the structure of LSTM Cell.

The gates are filters that determine what information is stored, read, written, and erased from the cells that share the same block. Each gate corresponds to a weights matrix; these can be learned when training the network. As a result, the information carried by these short-term components relative to the long-term components will progressively contribute less to the training of the model.

**Steps to forecast using LSTM**

1. Data Preprocessing: The input data are preprocessed to remove missing, redundant, and outlier values. Data preparation for the LSTM model includes the normalization of the dataset. In the current study min-max normalization to a [0, 1] range is used to create a lagged values data frame.
2. Train-Test split: The dataset is divided into training and testing sets. In the current study, 80% of the data is used for training and the remaining 20% are used for testing.
3. Build the model: Design the LSTM network architecture.
4. Train the model: Training involves making a prediction based on the current state, calculating the error, and updating the weights or parameters of the network to minimize the defined error. Thus improving the prediction accuracy.
5. Tune hyper-parameters using validation data.
6. Deploy the model for forecasting.

Design of LSTM-based model is for univariate time-series forecasting and multivariate time-series forecasting problems. Forecasting accuracy of the model can be improved by using the best features, optimal lags, defining layer, and training. The mean absolute error (MAE) and root mean square error (RMSE) are used to evaluate the accuracy of forecasting. Various adaptive learning optimization algorithms namely AdaGrad, RMSProp, and Adam are considered to improve the model performance.
4. Experimentation and Results

All experiments have been executed on a Core™ processor Intel® i7-4790 CPU 3.60 GHz with 16 GB RAM and Linux Ubuntu 14.04/ Windows 8 operating system. All codes are written in Python 2.7, using TensorFlow 2.0. Python Data Analysis Libraries used are NumPy, Pandas, Statsmodels and Matplotlib. NumPy and Pandas is a high level data processing and analysis library of Python.

These libraries provide an easy-to-use data structures. Pandas is built on functionality provided by the Numpy package and its key data structure is the DataFrame. Matplotlib is Python plotting library. Matplotlib is used to generate the output for the visualization[1]. Statsmodels provides statistical computations including descriptive statistics and estimation and inference for statistical models.

Keras deep learning library using the Tensorflow backend is used for Long Short Term Memory (LSTM) based prediction model [20].

A benchmark dataset “New Car Sales in Norway” is used for analyzing and predicting the car sales. The New Car Sales in Norway dataset is open source and downloadable from www.kaggle.com. The dataset contains the Monthly sales of new passenger cars by make (manufacturer brand), Monthly summary of top-20 most popular models (by make and model) and Monthly sales of new passenger cars by make (manufacturer brand) during 2007-2017.

Three separate experiment classes were carried out to Predict and Forecast Car sales using Kaggle Datasets of a Norway Company. The three classes are:

- Experiment C1: Exploratory data analytics is performed on dataset to summarize their main characteristics, discovering patterns and identifying anomalies with the help of summary statistics and graphical representations. Car Sales data have a lot of outliers and missing data. The outliers are cleaned and data is interpolated before using a time-series approach.
- Experiment C2: A single feature (Quantity) is used to make predictions for quantity of car sales in the future using New car sales by model univariate time-series dataset. Prediction of Car sales is performed using ARIMA and LSTM.
- Experiment C3: Prediction of Diesel and Green vehicle share is multivariate time-series prediction problem. The New car sales by month dataset contains more than one time-dependent variable. Diesel vehicle and Green vehicle share depends not only on its past values but also has some dependency on other attributes. Prediction of Diesel vehicle and Green vehicle share using ARIMA and LSTM is performed on multivariate time-series data.

Experiment C1: Exploratory Data analytics

Identify the top-10 manufacturers based on the total sale: In order to find top-10 car manufacturers, New car sales by make dataset is used. Firstly, calculated the total amount of the sales for each manufacturer from 2007 to 2017 and then sort them out. The experimental result shows the Volkswagen had been sold the most during the given period. The second popular brand is Toyota. Figure 2 shows the visual output of top 10 manufacturer based on total car sales in 10 years.
The sale of all the models of a particular manufacturer in a given year: New car sales by model dataset contains the sales figure of all the models in years between 2007 and 2017. The experiment has been carried out to find sales of various models of every manufacturer brand for every year. Figure 3 shows the results of sales of all the models of “Toyota” in year 2012.

Comparison of Car models on the basis of Percentage shares: New car sales by model dataset is used to find the share of all the models of all the manufacturer in Norway. Car selling quantity field is aggregated for all the models. Figure 4 shows the percentage shares of the car Models sorted in descending order.
The experimental result reveals the Volkswagen Golf having highest 6.23% percentage share whereas Mercedes-Benz CLA having the lowest 0.50% percentage shares.

Comparison of year-wise average consumption of CO₂ emission of all cars sold with year-wise average consumption of CO₂ emission in benzene-fueled cars sold and diesel-fueled cars sold: New car sales by month dataset has been used to carry out the experiment. The CO₂ emission of benzene-fueled cars, diesel-fueled cars and average CO₂ emission of all the cars are compared. Figure 5 shows a comparison of year-wise average consumption of CO₂ Emission of all cars sold. The results reveal that the average CO₂ is declining with the year progression. It is good for the improvement of the environment.
Comparison of sales of new, diesel and electric cars: Figure 6 shows the new car sales by month computed from sales data. Line chart is used to represent new car, diesel car and electric car sales.

![Sales of New Cars, Diesel Cars and Electric Cars during 2007-2017](image)

Figure 6: Sales of new cars, diesel cars and electric cars during 2007-2017

This can be observed from chart that the electric cars were introduced in January, 2011. The experiment reveals more sales of diesel cars as compared to electric cars. The line chart explicitly shows that the sales of electric cars are rising with time whereas the sales of diesel cars are dropping with time.

Experiment C2: Univariate time-series forecasting using ARIMA and LSTM

Predict the car sales for 2020 using the month-wise car sales quantity: The experiment uses New car sales by model dataset. Two models are implemented to observe the car sales trend in Norway, namely ARIMA and LSTM. The performances of these models are validated with the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The error measures are computed as follows:

\[
\text{MAE} = \sqrt{\frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N}} \tag{2}
\]

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{3}
\]

where \(\hat{y}_i\) is predicted value and \(y_i\) is actual value of N samples.

The drop in MAE and RMSE error rates for LSTM model as compared to ARIMA as depicted in Table 1, point towards the superiority of LSTM based forecasting.
Table 1: MAE and RMSE of ARIMA and LSTM models for Car Sale forecasting

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA</td>
<td>492</td>
<td>621</td>
</tr>
<tr>
<td>LSTM</td>
<td>333</td>
<td>422</td>
</tr>
</tbody>
</table>

Figure 7(a) shows the results when using ARIMA and Figure 7(b) when LSTM. The figure shows huge variation in residuals in case of forecasting using ARIMA. The result reveal that the model fitted with a supervised deep learning neural network known as LSTM-RNN is better time-series forecasting model than the ARIMA model.

Experiment C3: Multivariate time-series forecasting using ARIMA and LSTM

*Predict the diesel vehicles usage in future:* Conventional internal combustion engine vehicle running on petrol or diesel produces harmful impacts to the environment. It is suggested to reduce the use of such vehicles in order to make environment less polluted. This experiment is based on actual values found in *New car sales by month* dataset. Share of diesel vehicles depends on number of diesel-fueled cars sold; number of total cars sold and also on previous year share of diesel cars in total sales. A multivariate time-series forecasting is performed using ARIMA and LSTM based models. Table 2 summaries the performance of the models in terms of residual errors.

Table 2: MAE and RMSE of ARIMA and LSTM models for diesel vehicles share forecasting

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA</td>
<td>312</td>
<td>463</td>
</tr>
<tr>
<td>LSTM</td>
<td>251</td>
<td>339</td>
</tr>
</tbody>
</table>

Figure 8(a) shows the declination in share of diesel vehicles from year 2011 to 2017 and compares ARIMA prediction for year 2017 with actual share. Figure 8(b) validates LSTM prediction for year 2017.
Predict the green vehicles usage in future: A green vehicle is an eco-friendly vehicle that produces less harmful impacts to the environment than conventional internal combustion engine vehicles running on fuels such as, petrol or diesel. Green vehicles reduce harmful exhaust emissions and reduce oil imports in a country resulting into less pollution and fuel economy. With such intentions electric cars were introduced in Norway in Jan 2011. Here, multivariate time-series forecasting experiment is performed to predict the share of green vehicles in 2020. Sales of green vehicles is forecasted on the basis of sales of new hybrid cars, sales of new electric cars and number of used electric cars imported. The experiment results MAE and RMSE of ARIMA and LSTM models for Green vehicles forecasting is shown in Table 5.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA</td>
<td>292</td>
<td>428</td>
</tr>
<tr>
<td>LSTM</td>
<td>214</td>
<td>353</td>
</tr>
</tbody>
</table>

Figure 9(a) shows the growth of green vehicles from year 2011 to 2017 based on actual values found in New car sales by month dataset. ARIMA technique is used to compare the actual result from forecasted result and it is very well revealed that the figures are almost similar when performed on data up till Dec, 2015. Thus the same model is used to forecast the green vehicles usage for year 2019-2021. Figure 9(a) and 9(b) show the rise of green vehicles in years 2018, 2019 and 2020 using ARIMA and LSTM based forecasting model respectively.
5. Conclusion

The experiment of predictions of sales is performed by forecasting future trends in Norway's car sales data sets by utilizing the ARIMA and LSTM. The performances of these models are validated with the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

The result showed that the models fitted with a supervised deep learning neural network known as LSTM-RNN are found a much better time-series forecasting model than the ARIMA models. The experiment suggested future directions for the usage of deep learning-based models on predictive applications.

The study explored sales of used (import) cars, electric-used cars, and new diesel cars that suggest the sales of electric cars are rising with time whereas the sales of diesel cars are declining over time. The experiments identified the highest sold car and the most popular model of each manufacturer during the given period.

Comparison of Car models on the basis of Percentage shares suggested that Volkswagen Golf had the highest 6.23% percentage share whereas Mercedes-Benz CLA had the lowest 0.50% percentage share.

Comparison of CO₂ emission of benzene-fueled cars, diesel-fueled cars, and average CO₂ emission of all the cars reveals that the average CO₂ is declining with year progression. This is a great concern for the improvement of environmental protection.

The rise of green vehicles and the fall of diesel vehicle sales are also significantly drawn from predictive analytics performed using the models. The results elucidate that multivariate LSTM-RNN based time-series forecasting can be employed for effective and accurate forecasting in different domains such as credit, insurance, consumer behavior, and medical diagnosis.

References


