# **How A Large CPG Estimates Profitability Of Their Products On The Amazon.com Platform**

Akshay Madar, Ashish Tomar, Shivendra Kumar, Shubham Sharma, Sobhagya Singh Jamwal, Matthew A. Lanham

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

madar@purdue.edu; tomar0[@purdue.edu](mailto:Teammate1@purdue.edu); [kumar394@purdue.edu](mailto:Teammate2@purdue.edu); sharm405@purdue.edu; ssinghja@purdue.edu; [lanhamm@purdue.edu](mailto:lanhamm@purdue.edu)

## **ABSTRACT**

In collaboration with a multinational consumer products company that posts and sells its products on Amazon.com, this paper aims to provide the business an enhanced decision support system for identifying profitable as well as underperforming products, using both statistical and machine learning models.

The scope of this paper includes all vendors who sell their products on an ecommerce platform and find it challenging to understand the profitability of their product catalog, as compared to selling from a traditional brick-and-mortar retailer. Amazon shares mostly aggregated metrics and often is unwilling to provide more granular measures of where costs arose. For example, advertisement investment is not broken down by the product. There is a lack of clarity regarding how chargeback is performed and how fulfillment costs are derived. Large CPGs do not like working in the dark when the volume of products shipped and sold is massive, which leads to millions in costs.

This paper will help them to optimize product strategy and channelize the marketing efforts efficiently, via accurate sales forecasting and profitability prediction at the granular level of stock-keeping units (SKU).

**Keywords:** *CPG, SKU, ASIN, Time series forecasting, LSTM, Auto ARIMA, RMSE*

## **INTRODUCTION**

In today’s exceedingly competitive corporate setting, it is imperative that companies focus on profitability rather than merely on scaling up of revenue. Several aspects like inventory costs, marketing expenditures and product life cycle affect a business’s bottom line. Tracking the performance of each product in different product categories is a great potential solution to mitigate such problems.

Consumer product goods companies are increasingly moving their product portfolio to ecommerce platforms to tap into a massive online consumer base, with products spanning over multiple categories and price buckets. Among the various ecommerce ventures in online retail space, Amazon has emerged as one of the frontrunners. It provides the perfect stage to statistically gauge the different attributes associated with a product to maximize the profitability for the vendor, as compared to a brick and mortar store. The dynamics behind the association between Amazon and consumer product goods companies is complicated and convoluted under a highly complex web of data and algorithms.

This paper aims to unravel these intricate algorithms by developing a predictive model to estimate profitability of the products of the firm under scrutiny, listed on Amazon. The predictive results and business prescriptions of the model can be further extended to other ecommerce platforms.

The business problem is to use the key predictors, which determine profitability for their online business, to generate highly granular insights at the level of stock keeping units and different time horizons. These insights shall be used by the business on field to financially optimize their fund allocation, especially marketing ad allocation, for specific products of different categories. For example, if the reason for the falling sales of a signature product of the brand is found to be lack of online visibility to the customer, the company shall overcome this by improving its fund allocation. It shall provide the required marketing thrust to the online advertisement of that particular product.

Even the top global companies are highlighting these points in their reports. Accenture in its annual report underlined the importance of measuring the profitability of products, services and customer via a structured, methodical approach for appropriate cost allocations. PwC also suggested exploiting product portfolios for CPGs to drive profitability of SKUs. Intuit categorically highlights the role of SKUs to boost the profitability by providing better analysis of the stock and sales and improving the decision-making process.

Since, the cashflow pertaining to an online product depends on the number of units sold on Amazon, prediction of the number of units is integral for profitability determination. This paper tries to predict the number of products sold using historical data, in-house attributes, and other performance determining indicators like seasonality etc. These predictors are used to build a model which performs time series analysis over historical data to predict the future sales for a specific SKU. Certain parameters like chargebacks (penalty charged to the firm by Amazon caused by incompliance with shipping/fulfillment guidelines) have a considerable impact on the profitability identification.

Statistical models like Auto ARIMA and LSTM have been used to compare the accuracy of predicted sales, and the final model is selected based on higher accuracy Thus, the model results will assist the firm under scrutiny to optimize the overall profit by improved management of supply chain resources and optimization of the marketing expenditure on products at SKU level, based on its current level of profitability.

The remainder of this paper is organized as follows: A review on the literature on various criteria and methods used for supplier selection is presented in further sections. The proposed methodology is presented, the criteria formulation is discussed, and various models are formulated and tested. The paper then outlines the performance of our models, and with a discussion of the implications of this study, future research directions, and concluding remarks.

## **LITERATURE REVIEW**

Sales forecasting is an integral component of production and distribution planning, supply chain management, marketing campaigns which ultimately affects the profitability, especially in the consumer- packaged goods (CPG) industry.[1]

In the past decade, there have been several researches in the field of sales forecasting using several techniques. The literature review section has identified several papers related to application of machine learning and time series for sales forecasting. It established the supremacy of machine learning models over statistical methods. Most of the research work in the CPG industry has been done on holistic sales forecasting. Fig 1. shows the comparison of research in sales forecasting and profitability prediction for the CPG industry in the last decade. Research on sales forecasting and profitability prediction at SKU level is significantly less, as compared to research on total profitability for a firm. This paper takes it a level further by venturing into the dynamic space of online distribution, including not just the platform but also the external sellers. Amazon, being one of the biggest e-commerce companies, serves as an ideal platform to gauge the profitability prediction of the products at SKU level and forecast their sales.

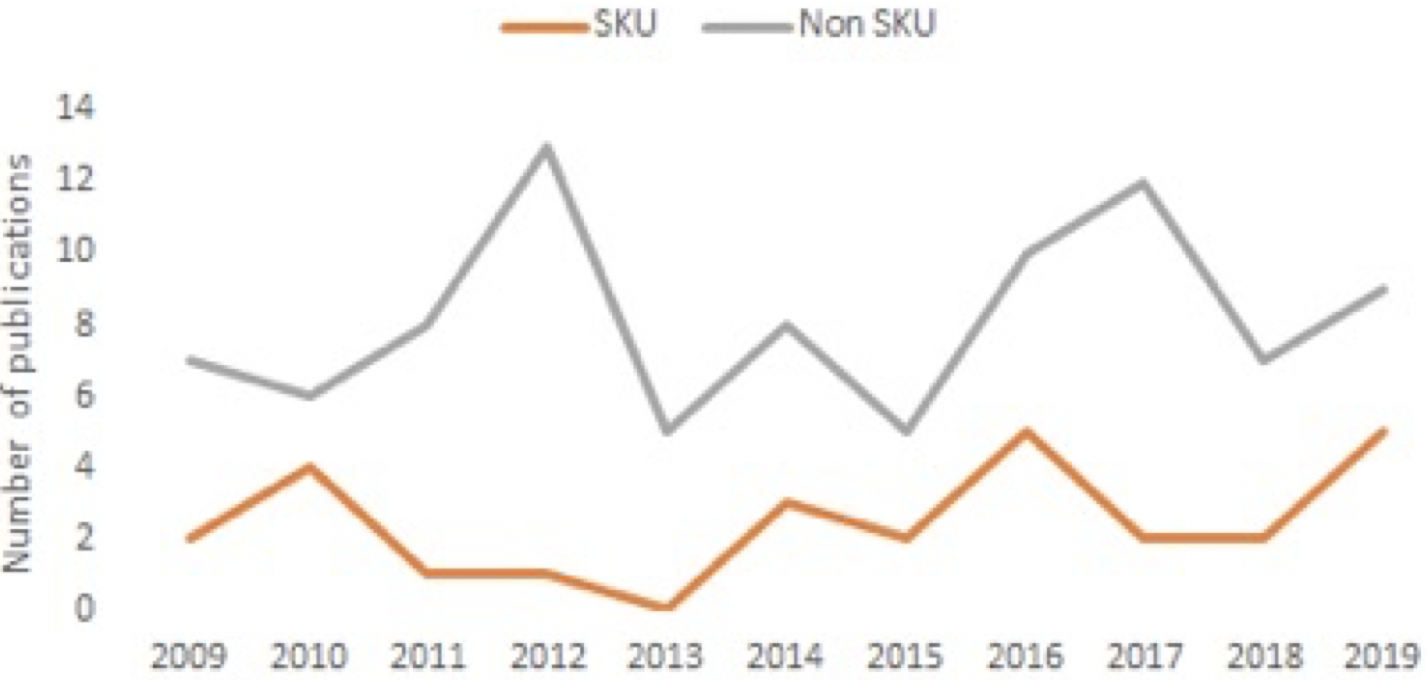


Fig.1 Research Trend

Forecasting accuracy can have a huge impact on overall monetary value of a regression model as minor changes in estimation of demand can cause drastic change in profitability. Furthermore, an under predicting or over predicting estimation can result in huge deviations from the real scenario. The prediction which underestimates sales can cause overstocking of inventory and the associated storage costs, while the one which overestimates sales can result in inventory getting out of stock and associated risk of customer dissatisfaction. Thus, this paper gives adequate importance to the accuracy of the model.

Methods like multivariate regression and KNN regression do not capture seasonality of the sales, which is a must for forecasting the sales based on time. Thus, time-series forecasting is employed. Normally, ARIMA models base the forecasting on only time as the independent variable. This paper also considers the impact of exogenous factors like sales price, price change, units on hand, weekly holidays, sentiment score through ARIMAX modeling to forecast sales.

Recurrent Neural Networks [2] has also been leveraged as a potential modeling technique to forecast sales and predict profitability with higher accuracy. It is a powerful neural network which is designed to study sequence dependence. The Long Short-Term Memory network or LSTM network is one of the prominent recurrent neural networks which can efficiently train large architectures. It is a highly effective deep learning model for time series prediction. The performance of the model is evaluated on hyperparameters like epochs, batch size, optimizer [3], activation function etc. In most of the practical applications, Adam works well as an optimizer function to give higher accuracy with little changes in hyperparameters. The data is scaled before fitting into the model to improve its convergence and learning rate. The result is further evaluated based on RMSE value.

While a lot of contemporary research on the subject fares well on the accuracy front, not much of it balances it out with the model cost. This paper gives special emphasis on the computational cost of the forecasting model by making appropriate improvements in the pre-modeling techniques. Although, individual time series models for all SKUs will capture the behavior of the product in a more granular fashion but will significantly increase the computational cost of modeling. Thus, this paper makes an attempt to optimize the modeling cost through two approaches- AutoARIMA and SKU segmentation.

The techniques for SKU segmentation can be broadly classified under two categories- 1) Judgmental, and 2) Statistical. Judgmental methods are driven by the opinions of managers while statistical methods use a wide variety of complex mathematical models to perform the grouping. This paper adopts the statistical approach of SKU classification because it provides a strong scientific reasoning for it. It performs clustering of the SKUs to group the similarly behaving products together, which not only keeps the model accuracy intact but also reduces overall model complexity.[4]

## **DATA**

Table 1. and Table 2. show the variables used in sales forecasting and profitability equation, their datatypes and informative descriptions.

**SALES FORECASTING**

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| WeekEnding | Datetime | Weeks from 2017-2020, containing sales data |
| RetailerSku | Categorical | Amazon Standard Identification Number (ASIN) - a 10-character alphanumeric unique identifier assigned by Amazon.com for product identification within the Amazon organization |
| UnitsSold | Numeric | Weekly number of units sold for that ASIN |

Table 1: Data dictionary for sales forecasting

**PROFITABILITY EQUATION**

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| OPS | Numeric | Ordered Product Sales |
| ASP | Numeric | Average Selling Price |
| COGS | Numeric | Cost of Goods Sold |
| PPP | Numeric | Pure Product Profit |
| PPM | Numeric | Pure Product Margin |
| Net PPM | Numeric | Net Pure Product Margin |
| VFCC | Numeric | Vendor Funding that offsets COGS |
| Quick Pay Discounts | Numeric | Discount given to Amazon for early invoice payment |
| Vendor Chargebacks | Numeric | Penalties charged by Amazon caused by non-compliance with shipping/fulfillment guidelines |
| Display Ad Revenue | Numeric | Revenue applied to Amazon Retail generated from Amazon Media (Search) Advertising |
| Freight Costs | Numeric | Costs to Amazon to transport goods from vendor FC to Amazon FC |
| Damage Allowance | Numeric | Costs to Amazon for customer returns, damaged items, concessions to customers etc. |
| Marketing Costs | Numeric | Costs to cover marketing teams at Amazon |
| Customer Shipping Costs | Numeric | Costs to Amazon to ship goods to customers |
| Fulfilment Center Fixed Costs | Numeric | Costs to Amazon for fixed warehouse expenses (Lease/Property Payments, Utilities, etc.) |
| Retail Team Fixed Costs | Numeric | Costs to Amazon to cover retail labor costs |

Table 2: Data dictionary for Profitability Equation

## **METHODOLOGY**

Fig 2. depicts the methodology workflow of the paper. The flowchart consists of 5 broad categories: Data, Data Preprocessing, Exploratory Data Analysis, Modeling and Results.

Data preprocessing was done on both the datasets- from Amazon and CPG partner. The dimensionality of the dataset was changed from long characteristic to short, to enable modeling on it. Missing values in the data were treated based on business acumen. Time series data of units sold was scaled using Min-Max scaling to suppress the effect of outliers and avoid updating of weights at irregular rate in the model.

As part of Exploratory Analysis, top 5 ASINs and the bottom 5 ASINs were established through the formulated profitability equation.

The model implementation was aimed to reduce the model complexity and hence the associated computational costs. There were 425 ASINs in total over a period of 3 years. 125 active ASINs were chosen which were currently active for further analysis.

The 3-year data of these active ASINSs was then divided into training set and test set. The initial 2 years of data was taken as training data, while the last year data was taken as the test data.

Since it is not computationally and financially possible for a firm to have models for each ASIN, there were two approaches adopted which would handle this- AutoARIMA and LSTM after clustering.

AutoARIMA was carried out on 105 ASINs and would select the best fitting model collectively for all the ASINs. On the other hand, LSTM, which would leverage neural network, required SKU segmentation. For this, clustering was carried out on the active ASINs, on the product type and the units ordered. Out of the 3 clusters obtained, 2 prominent clusters were chosen to proceed with modeling. Fig 3. shows the clustering performed on the active ASINs data.

Moreover, ARIMAX was also employed for model accuracy improvement by incorporating exogenous factors like sales price, price change, units on hand, weekly holidays, and sentiment score. But none of these external factors were found significant, thus no exogenous factor was considered for sales forecasting.

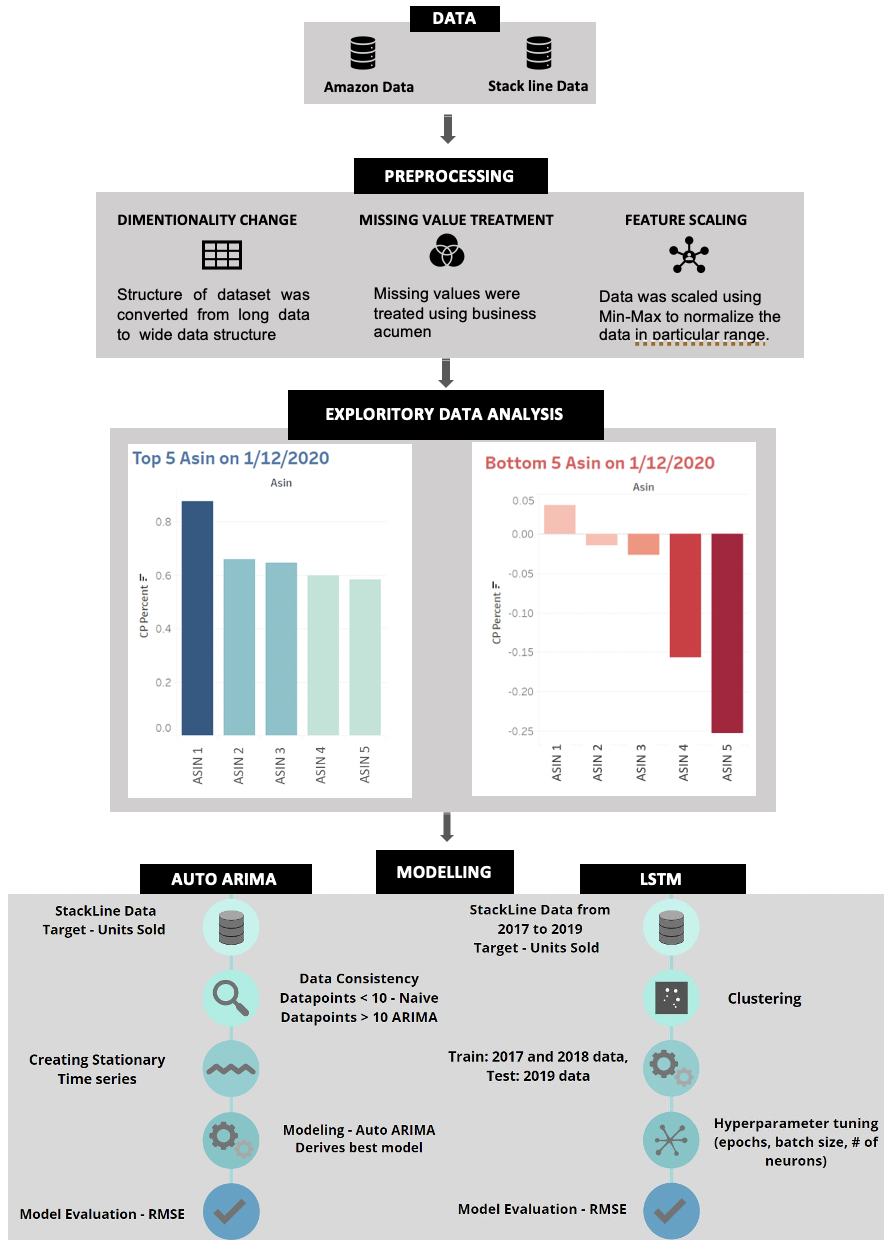


Fig 2. Methodology Workflow

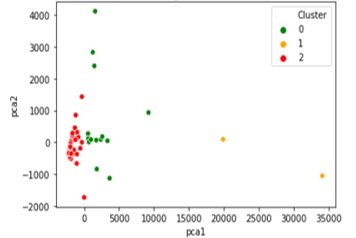


Fig 3. K-means Clustering on active ASINs

The accuracy of Auto ARIMA and LSTM models were then evaluated based on RMSE scores.

## **MODELING**

The nature of the underlying problem, which consisted of time-series data necessitated the use of the following models that could perform time series forecasting. The paper employs both, ARIMA forecasting algorithm and RNN based LSTM models for the purpose of the study.

LSTM:

Long Short Term Memory networks, usually called LSTMs, are a special kind of Recurrent Neural Networks, and have the capability to lean long term dependencies, which is crucial for time series forecasting. In addition to a single ‘tanh’ layer in conventional RNN, LSTMs have 4 additional interacting layers as shown in Fig X, which allows them to handle the ‘vanishing gradients’ problem.

On the flip side, LSTMs taken longer to run and occupy more memory to train. It is also vital to include dropout regularization to prevent overfitting, when using multiple LSTM layers.

For hyperparameter tuning, we controlled the following parameters – epochs, batch size, units, and dropout rate. [6]

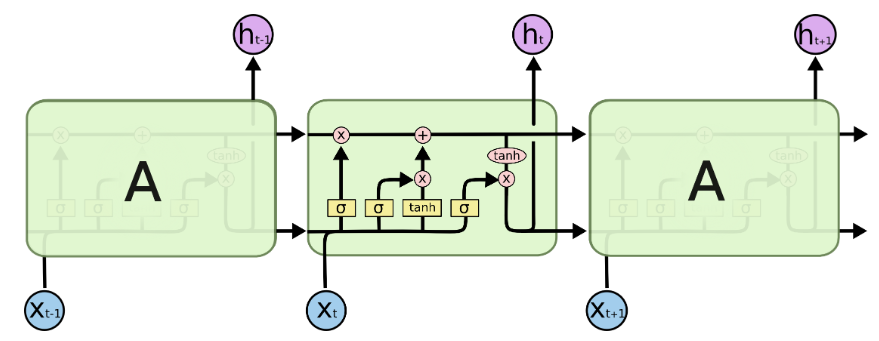


Fig. 4. LSTM Mechanism

ARIMA:

ARIMA or Auto Regressive Integrated Moving Average are a class of models used for forecasting time series data. The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

* p is the number of autoregressive terms,
* d is the number of nonseasonal differences needed for stationarity, and
* q is the number of lagged forecast errors in the prediction equation.

Auto ARIMA:

Pyramid, a python library brings R’s auto.arima functionality to Python. Unlike ARIMA, it uses brute force or stepwise approach [7] to try different combinations of p (auto-regressive), d (time backshift), q (moving average) and returns the best model after evaluation of AIC and BIC. Pyramid operates by wrapping [statsmodels.tsa.ARIMA](https://github.com/statsmodels/statsmodels/blob/master/statsmodels/tsa/arima_model.py) and [statsmodels.tsa.statespace.SARIMAX](https://github.com/statsmodels/statsmodels/blob/master/statsmodels/tsa/statespace/sarimax.py) into one estimator class, with a scikit-learn like user interface.

## **RESULTS**

The Fig 5. below shows the RMSE values for all the models covered in this paper on both training data and test data.

|  |  |  |
| --- | --- | --- |
| Model | AutoARIMA | LSTM |
| Train | 51 | 27 |
| Test | 63 | 38 |

Fig 5. RMSE scores of AutoARIMA and LSTM models

On comparing the RMSE values on the test data of all the models, LSTM was found to fit the actual values better with an RMSE score of 38 on test data. Auto ARIMA had an RMSE score of 63 on test data. Auto ARIMA shall be preferred over LSTM if ease of computation is the priority since clustering is not required. Also, it shall be used for long term predictions. LSTM shall be used for short term and accurate predictions.

Fig 6. And Fig 7. show the sales forecasting charts of AutoARIMA and LSTM respectively. The predicted units sold by LSTM evidently fits the closest to the actual values of the 3rd year.

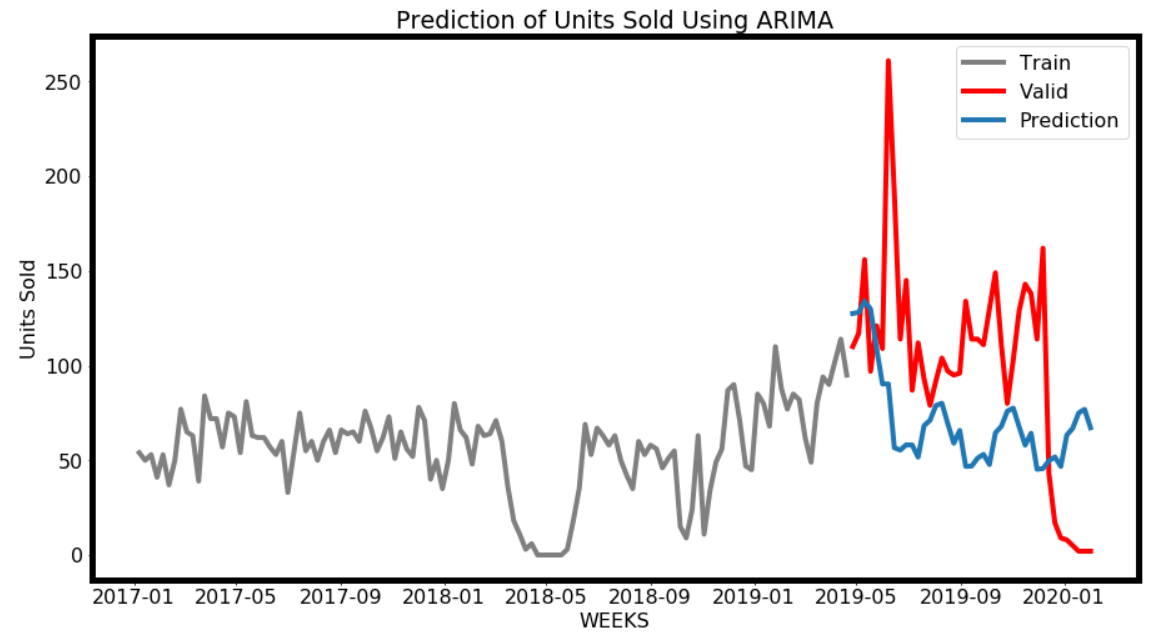


Fig 6. Sales forecasting by AutoARIMA

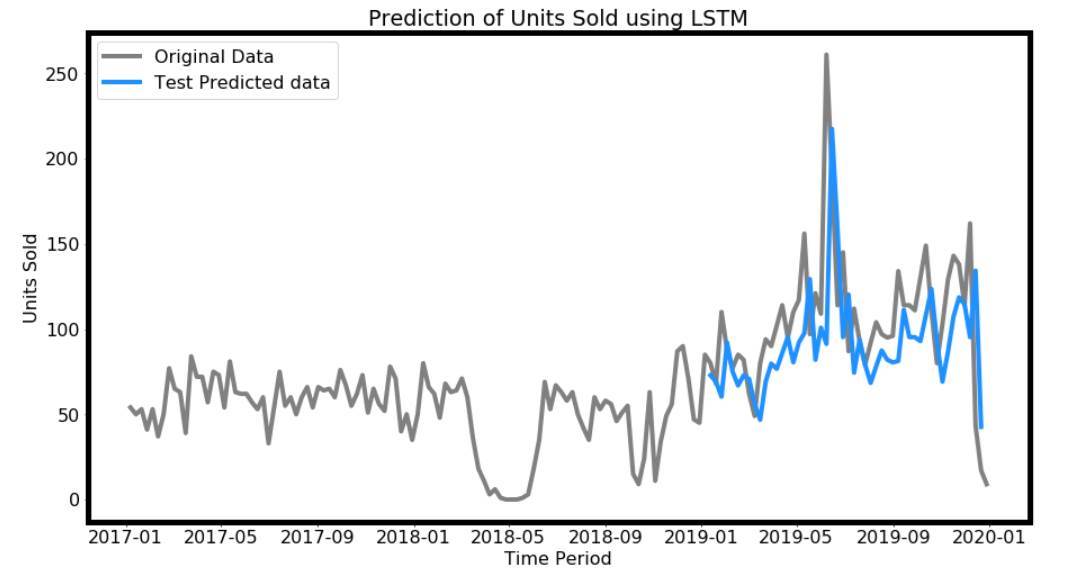


Fig 7. Sales forecasting by LSTM

## **CONCLUSIONS**

The business problem addressed in this paper was to predict the most and least profitable ASINs. Accurate sales forecasting for each product was also to be done, keeping the model complexity and associated computational cost as a crucial factor.

The paper identified top and bottom 5 ASINs from a profitability equation formulated by estimating the required Amazon metrics. ASIN drilldown approach was adopted to divide the number of ASINs according to the approach followed. Weekly sales forecast was achieved with high accuracy.

This paper assumes the ASINs in a particular cluster based on the product type and units sold, to behaver similarly. This assumption helps to achieve SKU level granularity by reducing computation cost of modeling. The future scope of the study could include incorporating exogenous factors such as competitor price and product attributes to improve the sales forecasting.

## **REFERENCES**

[1] Ramos, P., Santos, N., Rebelo, R., 2015. Performance of state space and arima models for consumer retail sales forecasting. Robotics and computer-integrated manufacturing 34, 151– 163

[2] Brownlee, J. (2019, August 5). Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras. Retrieved from https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/

[3] Eckhardt, K. (2018, November 29). Choosing the right Hyperparameters for a simple LSTM using Keras. Retrieved from <https://towardsdatascience.com/choosing-the-right-hyperparameters-for-a-simple-lstm-using-keras-f8e9ed76f046>

[4] Calster, T.V., Bossche, F.V., Baesens, B., & Lemahieu, W. (2020). Profit-oriented sales forecasting: a comparison of forecasting techniques from a business perspective. ArXiv, abs/2002.00949.

[5] Zhang, G. P., Qi, M., 2005. Neural network forecasting for seasonal and trend time series. European journal of operational research 160 (2), 501–514.

[6] Olah, C. (2015, August 27). Understanding LSTM Networks. Retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[7] Hyndman, R., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 27(3), 1 - 22. doi:<http://dx.doi.org/10.18637/jss.v027.i03>