

# SMU Data Science Review

---

Volume 3 | Number 1

Article 7

---

2020

## Demand Forecasting In Wholesale Alcohol Distribution: An Ensemble Approach

Tanvi Arora

*Southern Methodist University, tanvia@mail.smu.edu*

Rajat Chandna

*Southern Methodist University, rchandna@mail.smu.edu*

Stacy Conant

*Southern Methodist University, sconant@mail.smu.edu*

Bivin Sadler

*Southern Methodist University, bsadler@mail.smu.edu*

Robert Slater

*Southern Methodist University, rslater@mail.smu.edu*

Follow this and additional works at: <https://scholar.smu.edu/datasciencereview>



Part of the [Longitudinal Data Analysis and Time Series Commons](#), [Multivariate Analysis Commons](#), [Operations and Supply Chain Management Commons](#), [Statistical Methodology Commons](#), and the [Statistical Models Commons](#)

---

### Recommended Citation

Arora, Tanvi; Chandna, Rajat; Conant, Stacy; Sadler, Bivin; and Slater, Robert (2020) "Demand Forecasting In Wholesale Alcohol Distribution: An Ensemble Approach," *SMU Data Science Review*. Vol. 3 : No. 1 , Article 7.

Available at: <https://scholar.smu.edu/datasciencereview/vol3/iss1/7>

This Article is brought to you for free and open access by SMU Scholar. It has been accepted for inclusion in SMU Data Science Review by an authorized administrator of SMU Scholar. For more information, please visit <http://digitalrepository.smu.edu>.

# Demand Forecasting In Wholesale Alcohol Distribution: An Ensemble Approach

Tanvi Arora, Rajat Chandna, Stacy Conant,  
Dr. Bivin Sadler, and Dr. Robert Slater

Master of Science in Data Science, Southern Methodist  
University, Dallas TX 75275 USA

{tanvia, rchandna, sconant, rslater, bsadler}@mail.smu.edu

**Abstract.** In this paper, historical data from a wholesale alcoholic beverage distributor was used to forecast sales demand. Demand forecasting is a vital part of the sale and distribution of many goods. Accurate forecasting can be used to optimize inventory, improve cash flow, and enhance customer service. However, demand forecasting is a challenging task due to the many unknowns that can impact sales, such as the weather and the state of the economy. While many studies focus effort on modeling consumer demand and endpoint retail sales, this study focused on demand forecasting from the distributor perspective. An ensemble approach was applied using traditional statistical univariate time series models, multivariate models, and contemporary deep learning-based models. The final ensemble models for the most sold product and highest revenue grossing product were able to reduce sales forecasting error by nearly 50% and 33.5%, respectively, in comparison to a statistical naïve model. Additionally, this paper determined that there is no “one size fits all” demand model for all products sold by the distributor; each product needs an individually tuned model to meaningfully reduce error.

## 1 Introduction

In retail distribution and sales of beer and alcohol, demand forecasting is an integral and necessary part of supply chain management. A typical retail distribution is a three tier system as shown in Fig. 1. Ideally, a distributor should not only have enough inventory to satisfy customer demand but also be able to minimize the expense of purchasing and warehousing inventory [1]. Robust and accurate forecasting can positively affect many aspects of a company’s performance: it can assist in maximizing profits, managing product life cycles, and fostering collaboration between supplier and distributor [2, 3]. Conversely, poor demand forecasting can cause difficulties throughout the supply chain. Overstocks of inventory can incur excess costs caused by the necessity of surplus storage and stock expiration, while an under supply of stock results in lost sales and poor customer rapport [1]. Additionally, if the demand forecast is too low, suppliers may have to perform costly unscheduled production runs, which require additional labor for the supply chain [3].

Demand forecasting is a notoriously challenging task for any company. Beer and alcohol sales can be affected by myriad factors such as weather conditions, supplier's price promotions and price variations, holidays and sporting events, the state of the economy, and many more [3, 4]. Also, companies must invest time and resources into developing a forecasting technique and management style that best works for their specific needs and position in the supply chain. Management of demand forecasting includes making decisions about what information to gather, how to gather the information, who should be conducting the forecasting, what tools to use for forecasting, and what measures will be used to evaluate the forecasting [2].

Demand forecasting is, of course, not new to the alcohol distribution industry. However, in the past, forecasting has been complicated, slow, and dependent on a few employees with extensive domain expertise and experience [3]. The emergence of new forecasting methodologies has distributors and suppliers looking to advanced data analytics for more useful models. Many attempts at demand forecasting begin with traditional time series approaches that predict demand based on historical, sequential data points [5]. Popular time series approaches include autocorrelation models such as Autoregressive Integrated Moving Average (ARIMA) and multivariate models such as Vectorized Autoregressive model (VAR). The development of machine learning has advanced efforts concerning the demand prediction problem. Machine learning can harness big data that comprises more features and apply more advanced algorithms to make predictions. More recently, deep learning methods and neural networks have been applied to supply chain predictions, delivering improved results [6–8].

In this paper, historical data from a beer and alcohol distributor was used to forecast demand. As part of the initial literature review it was found that most of the work in alcohol demand forecasting is done from the retail point of view and not from the distributor's point of view. This work on alcohol demand forecasting for a distributor is a novel contribution to this domain. The novel approach to this problem was an ensemble model of traditional univariate time series models, multivariate models, and deep learning-based models. The overall goal was to create more accurate forecasting in order to optimize inventory, improve cash flow, and enhance customer service for the distributor.

The remainder of this paper is organized as follows: In Section 2, related academic works are reviewed. In Section 3, the details of the methodologies are presented. Intermediate models and the final ensemble methods are also discussed in Section 3. Section 4 describes the Methods of Evaluation, based on which the intermediate and final models were judged. Data and exploratory data analysis (EDA) are considered in Section 5. It contains information about the dataset the initial analysis that led to some key decisions on the overall approach to the problem. Decisions made in Section 5 lead to the Experimental Setup detailed in Section 6. Section 7 provides the results and performance analysis of all the models for an alcohol distributor's most sold product and highest grossing product. Lastly, in Section 8, this work concludes with consideration for future related studies.

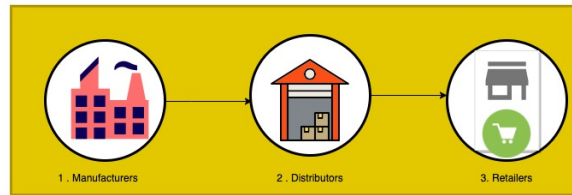


Fig. 1: The Three Tier Distribution System

## 2 Related Work

Traditionally, demand forecasting was dominated by time series and linear methods as they were well understood and effective on most of the problems. Deep learning models are black box but are able to learn arbitrary complex mappings from inputs to outputs automatically. In this paper, an ensemble of best of traditional models and newer deep learning models were explored and created.

Dejan Mircetic et al., in 2016, [9] created several forecasting models of beverage consumption for one of the market leaders in South-East Europe. They used an S-ARIMA model to forecast consumer beverage demand patterns because of its flexibility and ease of understanding. Root means square error (RMSE), mean absolute percentage error (MAPE), mean absolute scaled error (MASE), among others, were used as the evaluation criteria for comparing multiple models. Their study found that models that included only auto regressive and moving average components were more successful than models that also included seasonal differencing. Surprisingly, they also found that the simple average naive forecast model outperformed S-ARIMA models that included seasonal components. They concluded that this result should be tested on other beverage companies to determine if this was particular to the company in their study, or whether it could be generalized to the entire industry. While Mircetic et al. provided a useful outline for testing and evaluating S-ARIMA in the beverage supply chain, their focus was on consumer demand instead of distributor demand, and a different result could be expected for the forecasting models in this paper.

Artificial Neural Networks have been around for a long time, but with the advent of faster hardware and highly optimized open-source libraries, broad and deep neural networks have impressively and skillfully been applied to a range of problems. Deep Learning methods offer promise for time series forecasting as well, and much recent work in this area has been done on Deep Learning models.

Xinyu Hu et al. in Jan 2019 [10] proposed a regression-based Bayesian Neural Network (BNN) model to predict spatiotemporal quantities, such as hourly rider demand, with calibrated uncertainties. The paper has two contributions: (i) A feed-forward deterministic neural network architecture that predicts cyclical time series data with sensitivity to anomalous forecasting events; (ii) A Bayesian framework applying Stein variational gradient descent (SVGD) to train large neural networks capable of producing time series predictions, as well as, a mea-

sure of uncertainty surrounding the predictions. This BNN reduced average estimation error by 10% across eight U.S. cities compared to a multilayer perceptron (MLP), and 4% better than the same network trained without SVGD. Remaining interesting research questions explored different correlation structures to model time series data and investigation of the use of more structured prior distributions instead of prior independent assumptions made in the paper. This use case is similar to the distribution demand predictions that this paper will explore, which involve multivariate time series involving product, time, and location.

Predicting financial time series is another common use case in this problem area. Many researchers have used deep learning methods in recent years. Sangyeon Kim and Myungjoo Kang in March 2019 [11] compared various deep learning models such as MLP, one-dimensional convolutional neural networks (1D CNN), stacked long short-term memory network (LSTM), attention networks and weighted attention networks for financial time series prediction. Attention LSTM can be used not only to predict but also to visualize intermediate outputs for analysis on the mechanics of the prediction. It can help to understand why certain trends are predicted when accessing a given time series table. Their model produced a 0.76 hit ratio, which was found to be superior to those of other methods. Their experiments considering 60 trading days as lookback days and 40 trading days as prediction days returned the highest hit ratio with the attention networks model. Moreover, the highest earn points were returned with weighted attention networks, as loss functions were minimized when improved at higher change ratios. These performed better with long sequential data. This paper applies some of these advanced methods to the field of demand forecasting in alcohol distribution.

Some researchers have successfully used hybrid or ensemble approaches that marry traditional time series methods with machine learning and deep learning [6–8,12]. A study conducted by Islek and Oguducu in 2015 [7] sought to reduce demand forecasting error by clustering warehouses and sub-warehouses for a Turkish dried fruit and nut company according to their sales behavior. They then used a hybrid forecasting approach that combined a time series moving average model and a Bayesian Network machine learning algorithm and used separate models for each cluster. When using one model for all warehouses, the result was a 49% Mean Absolute Percentage Error (MAPE). After applying their specific models to each warehouse cluster, the researchers achieved 17% MAPE.

In 2019, Kilimci et al. [6] also used an ensemble approach for demand forecasting, this time for a Turkish retail grocery store. They integrated nine different traditional time series methods, a support vector regression (SVR) algorithm, and a deep learning multilayer feed-forward artificial neural network (MLFANN). The authors also created an integration strategy to combine the strengths of the different approaches into a “single collaborated method philosophy”. The integration strategy considered weekly model performance and gave weight to predictions from the more accurate models for that week. They conducted forecasting on multiple product groups and found a reduction in MAPE for all groups when using their novel integration strategy, and further improve-

ment was seen when their deep learning method was also employed with the integration strategy. Like these examples, this paper will also utilize an ensemble approach, but will novelly apply it to the distribution of wholesale alcohol using multiple traditional times series models and LSTM models.

### 3 Methodologies

Below are the models that are analyzed and used for this paper's solution:

**S-ARIMA:** Autoregressive Integrated Moving Average (ARIMA) is one of the most widely used forecasting methods for univariate time series forecasting. ARIMA uses its own past values of the time series and has a treatment for the random factors through the use of moving averages [8, 12, 13]. However, it has certain limitations when it comes to seasonal data, i.e., a time series with a repeating cycle. Seasonal Autoregressive Integrated Moving Average (S-ARIMA or Seasonal ARIMA) is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. An S-ARIMA model is formed by including additional seasonal terms in the ARIMA. The seasonal part of the model consists of terms that are very similar to the non-seasonal components of the model, but involve back shifts of the seasonal period [9].

**Vector Auto-regression (VAR):** VAR is a very useful statistical model that can capture linear inter-dependencies among multiple variables, or time series [14]. In VAR, each time series is modeled as a function of past values and treats all variables as dependent. Unlike univariate models, VAR models are bi-directional, meaning that the model takes into account how variables influence each other. This makes VAR a standard means of modeling and forecasting in many fields, such as economics, that have a need to assess the subtleties and interrelationships between the variables of interest [15, 16].

**Long Short-Term Memory Networks (LSTM):** Recurrent neural networks (RNN), like the Long Short-Term Memory network, or LSTM, can explicitly handle the order between observations when learning a mapping function from inputs to outputs. This works very well with sequence data, which is the case for time series input [17]. Where standard RNNs fail to learn in the presence of time lags greater than 5-10 discrete time steps between relevant input events and target signals, LSTM can learn to bridge minimal time lags in excess of 1000 discrete time steps by enforcing constant error through "constant error carousels" (CECs) within special units, called cells. LSTMs solve both the problems of vanishing gradient and exploding gradient. A basic LSTM Network architecture is shown in Fig. 2 [18].

The first step in LSTM is the "forget gate layer", that decides what information will be thrown away from the cell state. It looks at  $h_{t-1}$  and  $x_t$ , and outputs a 0 representing "get rid of this" and a 1 representing "keep this". Next is the "input gate layer", which decides which values will be updated. A  $\tanh$  layer creates a vector of new candidate values,  $C_t^-$ , that could be added to the step. These two are combined to create an update to the state. New Cell state  $C_t$  is created by multiplying the old state, i.e.  $C_{t-1}$ , by the forget gate layer  $f_t$  and

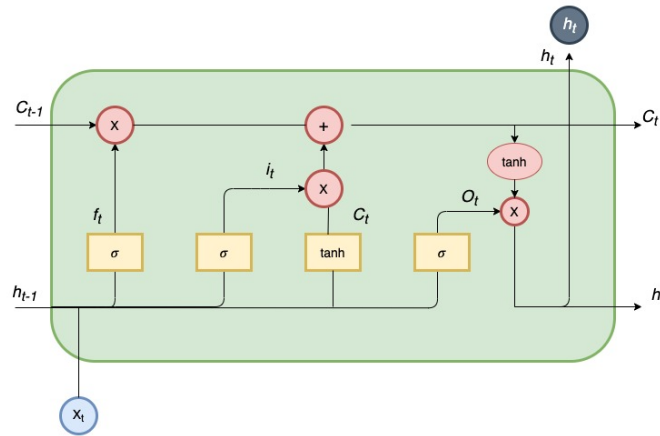


Fig. 2: Basic LSTM Architecture

puts the cell state through **tanh**,  $C_t$ , so that it would output only the parts that were desired.

**Ensemble Models :** Ensemble modeling is a process where multiple diverse models are created to predict an outcome and then aggregated to get a final prediction. This helps to improve overall performance and reduce biases [6]. Modes of aggregation can be either weighted i.e., different weights are assigned to different models based on their importance, or non-weighted i.e., equal weight is given to all models. For the dataset used in this paper on alcoholic beverage sales, a custom ensemble model based on different products was found to be the most useful.

## 4 Methods of Evaluation

Different evaluation criteria were used during different phases of model building to evaluate the model performance. To evaluate the various created models as candidates for the best identified model, for each modeling approach - univariate, multivariate and deep learning based - the following criteria were used:

**Akaike Information Criterion (AIC):** AIC is an estimator of overall model quality and widely used for model selection. It measures the relative information loss by a given model. Thus, less information loss by a model, better the model quality [19].

$$\mathbf{AIC} = 2k - 2\ln(\hat{L}) \quad (1)$$

**Bayesian Information Criterion (BIC):** BIC is closely related to AIC, but imposes a larger penalty for a higher number of parameters in a model as compared to AIC. Thus, models with lower BIC scores are preferred [19].

$$\mathbf{BIC} = \ln(n)k - 2\ln(\widehat{L}) \quad (2)$$

where

- $\widehat{L}$  = the maximized value of the likelihood function of the model  $M$ , i.e.  $\widehat{L} = p(x | \widehat{\theta}, M)$ ,  $\widehat{\theta}$  are the parameter values that maximize the likelihood function;
- $x$  = the observed data;
- $n$  = the number of data points in  $x$ , the number of observations or equivalently, the sample size;
- $k$  = the number of parameters estimated by the model. For example, in multiple linear regression, the estimated parameters are the intercept, the  $q$  slope parameters, and the constant variance of the error: thus,  $k = q + 2$ .

**Forecasting errors** were evaluated in terms of Average Square Error (ASE) or Root Mean Squared Error (RMSE). Lower ASE or RMSE yielding models are preferred [20].

$$ASE = \text{mean}((ForecastValue - ObservedActualValue)^2) \quad (3)$$

$$RMSE = \sqrt{ASE} \quad (4)$$

**Quality of Residuals:** The models that tend to whiten the residuals the most are considered to model the trend and pattern in the dataset. Hence, models that whiten the residuals to a higher degree are preferred.

**Forecasting Characteristics:** Some models perform better on short term forecasts, whereas other models perform better for long term forecasts. Hence, the forecast duration also plays a vital role in the model evaluation phase. Preferred models would be those that perform better with forecast duration specified by the distributor.

For evaluating the candidate models to find the best identified model for each of the approaches in the ensemble, both **forecasting errors**, in terms of ASE and **forecasting characteristics** produced from the model, were used in a sequenced cross-validation manner to obtain the best fitting final model.

## 5 Dataset

This section provides details on the dataset that was used for this study and preliminary analysis on it.

The data set for this study was from a large American wholesale alcohol distribution company (Company A), which deals in beer, spirits, wine, and a



few other alcohols. Initial data contained the total number of standard cases sold per month for different beverage types “spirit” at a distribution warehouse located in South Texas. There were different products with varying quantities for each beverage category in the dataset. The goal was to predict, given the current dataset, the number of standard cases for six months horizon into the future, and the predictions should be updated monthly.

As the testing progressed, several other factors that could affect the demand for alcohol in a month were considered and added to the dataset. These additional predictors were:

- Local weather
- Price promotions
- Marketing campaigns
- Holidays and special events (festivals, sports, concerts, etc.)

### 5.1 Exploratory Data Analysis

This subsection provides details on exploratory data analysis performed on the dataset used for this study.

The sales and pricing data provided by Company A was for seven years (84 months), from January 2013 to December 2019. It included the target variable: the number of standard cases ordered per month by different customers for a certain product ID. Each product ID is unique for a product of a particular size. There were no missing values in the received dataset. However, a close examination revealed that if there was no demand for a product in a particular month, then data for that product is not recorded for that particular month.

The distributor provided a list of attributes and metrics contained in the dataset. The primary focus was to predict the demand for standard cases. Additional sales information that the client maintains and that was not shared was Makeup Cases, House Goals, Free Goods, and Credit cases. All of these factors could impact the demand or could themselves be by-products of previous sales.

The dataset received contained 375K observations and 29 features for monthly sales and pricing data. Company A has seven distributors in Texas alone, distributing four different types of beverages, including beer, wine, spirits, and others. The alcoholic beverage considered for this study was spirits. Spirits alone had 54 different types of products, such as whiskey, rum, vodka, etc. which had a total of 405 different brands. These brands include Jack Daniel’s, 1800 Tequila, 360 Vodka, etc. All of these brands combined to 4017 unique products. Each product came in different sizes like 1L, 200M, 750M, etc. There were a total of seven sizes available. Orders were maintained as cases where a case is a unit of measure equal to a case of beer or 24 12-ounce bottles (2.25 gallons).

Based on the EDA conducted for all the spirits in the given dataset, Taaka Vodka 80 1L was found to be the most sold product in the last seven years and Jack Daniel’s Black Whiskey 1L was found to be the highest revenue grossing product for the last seven years. The prediction analysis was focused on these two products, since forecasting these products accurately would benefit distribution

house the most. Vodka and Scotch are two very different products with different flavors and customer appeal. The times series for cases sold for Taaka Vodka (Fig. 3) and Jack Daniel's (Fig. 4) were very different so it was expected that there would be varying factors that affect the orders from year to year; and that one demand model may not generalize to both products.

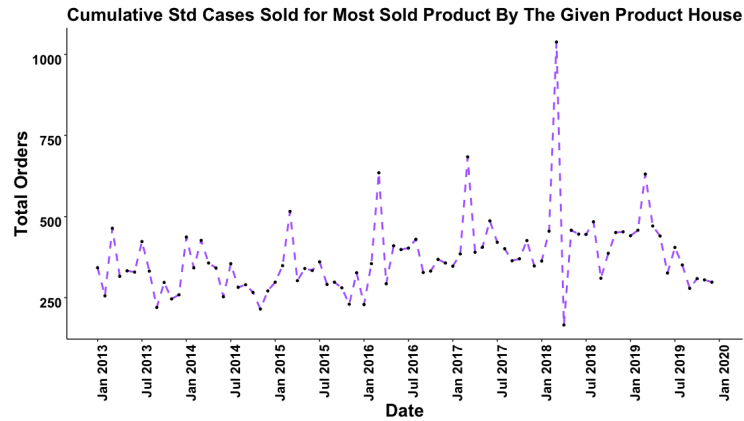


Fig. 3: Taaka Vodka 80 1L - Total Cases Sold Per Month

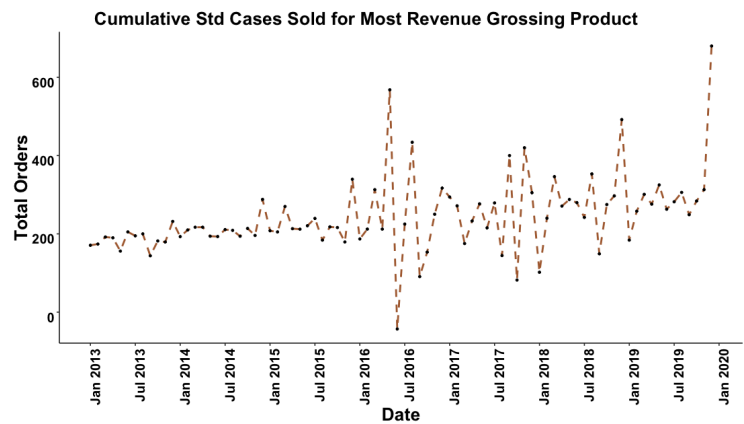


Fig. 4: Jack Daniel's Black Whiskey 1L - Total Cases Sold Per Month

## 6 Experimental Setup

**Individual models for a particular product ID.** This study began by developing a proof-of-concept model for demand forecasting for the distribution center’s best selling product: Taaka Vodka 80 1L. Univariate, multivariate, and deep learning models were created and a rolling window cross-validation strategy was used to identify the best models.

The rolling window cross-validation is a technique that implements cross-validation while accounting for the sequenced nature of the time series [21, 22]. Rolling windows, a variation of back testing, also called sliding windows, have long been used in forecasting financial data. Elements of a rolling window model include the window size, the horizon, and the rolling steps. Fig. 5 graphically explains this logic. The window size is the number of data points in the training set; for this research, 36 months of past standard case sales data was used. The horizon is the forecasting window size for unseen data, which was chosen to be six months for this study. The rolling steps are the number of data points skipped on each pass [21]. The step size applied was six so that forecasts at each step do not overlap. Each model’s performance was judged on average of ASEs from the eight, six month sliding windows.

**Create ensemble model.** Once best models from the individual model types i.e., Univariate, multivariate, and deep learning models, were identified, they were combined into a simple average or weighted average ensemble model, depending upon the relative credibility of obtained models.

**Feasibility study.** Next, a feasibility study was conducted to investigate whether the developed hybrid model on Taaka Vodka could be generalized to different products of the same type “spirit”, but different beverage types. The other product chosen was Jack Daniel’s Black 1L, the highest-grossing product for the distribution center in question.

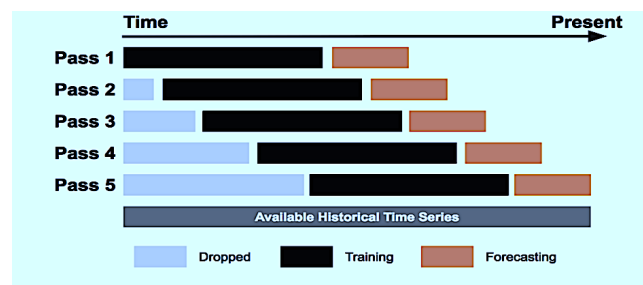


Fig. 5: Rolling Window: Cross Validation Strategy For Model Selection

## 7 Results and Analysis

This section details the results that were obtained after fitting the best identified models on the distribution house's most sold product, Taaka Vodka, and highest revenue grossing product, Jack Daniel's Whiskey, and the subsequent analysis that was performed on the obtained results.

### 7.1 Taaka Vodka

This subsection details the results and the subsequent analysis that were performed for the distribution house's most sold product: Taaka Vodka. Specifically, all of the model fitting began with total cases of the Taaka Vodka 80 1L product.

**Naive Model:** As a benchmark to judge the more sophisticated models, a statistical naive model was created. As per this model, predictions for a given month of a year were calculated as a simple average of the same month for the previous three years. The average ASE of all rolling windows for Taaka Vodka was found to be 12631.85. Using the naive ASE as a benchmark allowed a better understanding of the value that was added by the more intelligent models being tested later. Fig. 6 displays the forecast for the last four years with the actual sales.

**S-ARIMA:** The S-ARIMA model sought to improve upon the Naive Model by exploiting any autocorrelation structure that might be present in the data. For model identification, after making the time series stationary by removing  $s=12$  seasonality from it, a total of 114 univariate models were created and tested using grid search with order of auto-regressive,  $AR(p)$  terms ranging from 0 to 18 and order of moving average terms,  $MA(q)$  terms ranging from 0 to 5. Two additional models were created by first order and second order differencing the time series with removed seasonality. Both of these models were not found to be useful when compared to models with only seasonality removed. Out of all 116 models attempted, the  $ARIMA(12, 0, 0)$  with a seasonality of 12 was found to have the smallest ASE of 7619.15, a 40% improvement from the Naive Model. It was not surprising that the data showed a monthly seasonality as the number of cases per month is often similar to that of the same month in previous years. Fig. 7 displays the forecast of the last four years with the actual sales using the univariate model.

**Vector Auto-Regression (VAR):** The VAR modeling began with a bivariate approach using the number of cases sold and the average price per case for the Taaka Vodka with various seasonality and lag. Then, to improve forecasting further, weather data was added to the model. Weather conditions have been found to influence retail sales and the type of goods purchased by consumers [4, 23, 24]. This weather data was obtained for the distribution center's region from the National Oceanic and Atmospheric Administration's (NOAA) Climate Data Online service [25]. After cleaning the data, the average monthly temperature and average monthly rainfall were added to the monthly time series. However, it was found out that, for this product, the addition of the weather data lead to increased noise in predictions, and hence, these weather predictors were

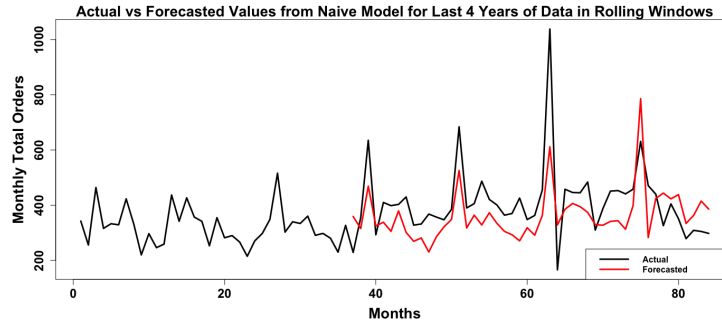


Fig. 6: Naive Model - Forecast of Taaka Vodka Total Cases per Month

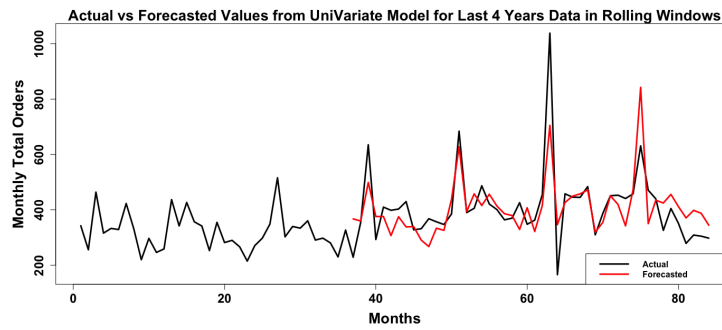


Fig. 7: Univariate Model - Forecast of Taaka Vodka Total Cases per Month

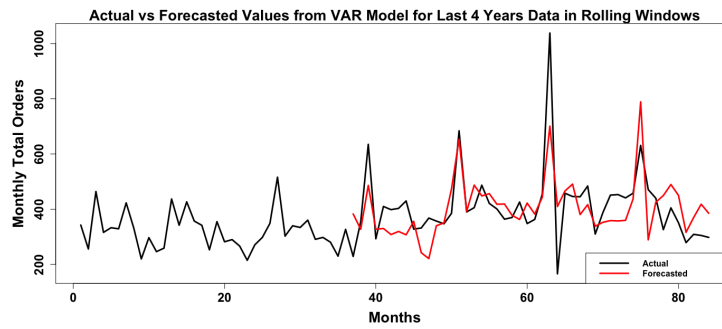


Fig. 8: VAR Model - Forecast of Taaka Vodka Total Cases per Month

subsequently removed. A total of 36 models, with seasonality = 12 accounted for, were tested in grid search fashion with parameter for highest lag order selection according to the AIC score, ranging from 0 to 3. The grid search also include the parameter for which type of deterministic regressors to include: 'none', 'const', or 'both). The features included in the grid search were sales, unit price, average temperature and average rainfall. The final multivariate model, a VAR(lag=3, sales, unit price), displayed monthly seasonality and attained an ASE of 9852.90. This was a 22% improvement over the Naive Model. Fig. 8 displays the forecast for the last four years with the actual sales using the VAR model.

**Long Short-Term Memory Networks (LSTM):** LSTMs, amongst the deep learning models, are known to be well-suited for making predictions based on time-series data. Each of the LSTM models was run on 36 different lookback window sizes, or lag, i.e., how many previous inputs were used to predict the next value. The initial approach was to apply a base LSTM model to the time series data in a univariate fashion. The best rolling window ASE obtained was 4670.468 for a look back size of 10.

As part of this study, the LSTM model was enhanced by adding other predictors as auxiliary inputs to the model. Based on the data received from Company A, price was the only other predictor that could be used as an auxiliary input. It was also identified during the EDA that impending price changes could affect product sales. Because of this identified lagged relationship between sales and average product price, in which price changes in immediate past influence future sales, models with price difference instead of contemporary price performed better. During the literature review, it was found that holidays or football season could impact sales of alcohol. Company A did not supply this data, so additional variables were created to define the number of national holidays in a month; and the months that are the primary season for NFL and college football. These additional predictors were appended to the monthly sales data, and new models were created by adding a combination of these as auxiliary inputs. Auxiliary inputs were added as linear sequences to the output of LSTM. Based on the tests conducted, the model with auxiliary input from the last lagged input i.e., from  $t-1$ , gave the minimum rolling window ASE. This model took an input of sales, the size of which was based on the lookback value, and passed through an LSTM layer with 32 hidden units. If there was an auxiliary input(s), it was concatenated with the LSTM output and then passed through 3 Dense layers with 64 hidden units and activation function as 'relu'. A final output layer was added with the number of outputs as six, representing the next six months' predictions and an activation function of 'sigmoid'. The model was compiled using `mean_squared_error` as the loss function and 'adam' optimizer.

The best LSTM model used impending price change and achieved an ASE of 2960.519, a significant 76% improvement over the Naive Model ASE. Table 1 provides the summary of LSTM models created with corresponding ASEs obtained for the best model and Fig. 9 displays the forecast for Taaka Vodka using the model with the lowest ASE (`lstm_pricediff_auxinput`), which had the smallest ASE amongst the LSTM models tested.

Table 1: ASE of best models amongst different LSTM models for Taaka Vodka

LSTM Models	# of Models	ASE of Best Model
tuned_base_lstm	36	4670.468
lstm_with_price	72	7309.540
lstm_with_price_hol_games	108	6668.175
lstm_with_pricediff	36	2960.519
lstm_with_pricediff_avgtemp_avgRain	36	6934.131

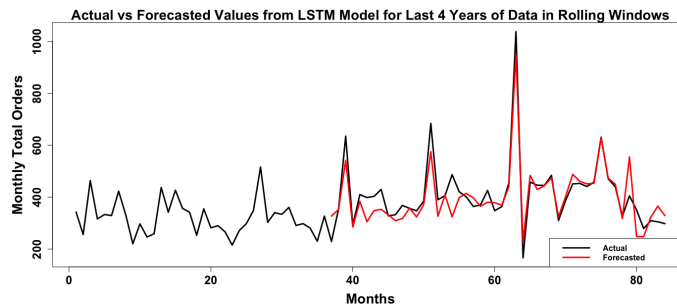


Fig. 9: LSTM Model - Forecast of Taaka Vodka Total Cases per Month

**Ensemble:** Finally, an ensemble demand forecasting model was created integrating best identified univariate, multivariate, and LSTM models. An ensemble model could avoid possible over or under-fitting present in its constituent models. Any shortcomings of one model would be compensated by the rest of the models in the ensemble. For Taaka Vodka, a weighted ensemble was employed that combined the models in ratio to their accuracy. Weighting the ensemble allows the more accurate model (the LSTM in this case) to have more importance in prediction than a less accurate model. To create the weighted model, the mean ASE of the three models were summed to create the total error. This total error was used to weight each model individually using their mean ASE. The weight of the model was computed as ASE of the model divided by total ASE from all the models in the ensemble. These weights were then multiplied by their corresponding model forecasts, and the result was summed. Lastly, this total was divided by the sum of the weighted ASEs. The weighted ensemble model achieved an average ASE over all rolling windows of 3633.82, a 71% improvement over the Naive Model ASE. A graph of the ensemble forecast and the actual monthly cases of Taaka Vodka sold is displayed in Fig. 10. A simple average ensemble model was also created but weighted ensemble model was finally chosen since it gave more weight to predictions from most accurate constituent model.

Table 2 provides a summary of the total number of models created and tested for Taaka Vodka and the best-identified model with the smallest ASE for each of the model types.

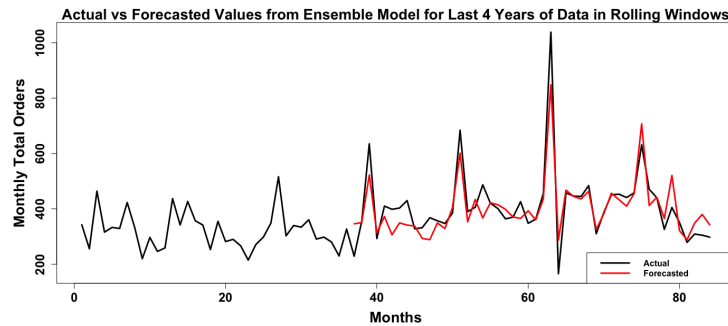


Fig. 10: Ensemble Model - Forecast of Taaka Vodka Total Cases per Month

Table 2: Summary of demand models tested and best ASE for Taaka Vodka

Model Type	No. of Models Created and Tested	Best Identified Model ASE
Intuition based Naïve Model	1	12631.85
Univariate Model	116	7619.15
VAR Model, Multivariate	36	9852.90
LSTM	288	2960.52
Ensemble	2	3633.82

## 7.2 Jack Daniel's Whiskey

This subsection details the results and the subsequent analysis performed for the distribution house's highest revenue grossing product: Jack Daniel's Whiskey. Some of the experiences gained from the Taaka Vodka analysis were considered for prediction models for Jack Daniel's. Hence a fewer number of models were developed and tested.

**Naïve Model:** As was done with the Taaka Vodka, a statistical naïve model was created based on the averages of the same month from the previous three years. The average ASE of the all rolling windows was 14521.89. Fig. 11 displays the forecast for the last four years in rolling windows with the actual sales using the Naïve Model.

**S-ARIMA:** A univariate model was explored, but the time series had inconsistent variance that resembled a white noise pattern. It was decided to split the time series into two parts and then model each part as a separate process. However, even after splitting the series at the point of change in variance, the two obtained sub time series still resembled white noise pattern. It was decided that a univariate model would not be an appropriate choice for this time series.

**Vector Auto-Regression (VAR):** For Jack Daniel's Whiskey, 52 different VAR models were explored manually by adding and removing predictors, starting with the following predictors included in the model: the number of cases sold, the average price per case, the average monthly temperature and rainfall for the area



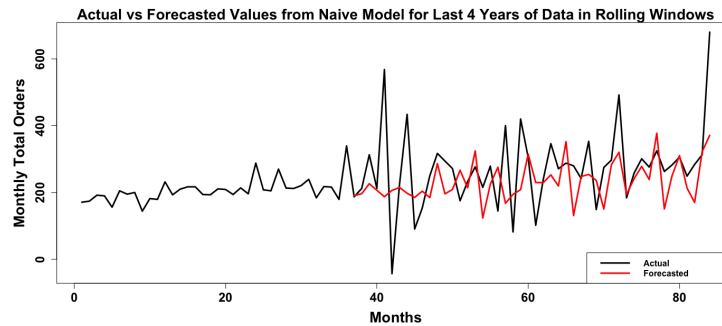


Fig. 11: Naive Model - Forecast of Jack Daniel's Total Cases per Month

with no seasonality parameter and highest lag order set to 1. Time was also added as an exogenous variable. Later, 2, 3 and 4 way interactions among these variables with different lag values were also added into the models. This was done to increase the complexity in the models as base VAR models, without interaction terms, demonstrated high bias. The best identified VAR model included the following variables as suitable predictors: Sales, Average Price, Average Rain, Average Temperature, Exogenous time variable, the interaction between rain and time, the interaction between price, temperature, and time; and interaction between price, rain and time. This final multivariate model achieved an ASE of 14402.26. This was only a 1% improvement over the Naive Model for the Jack Daniel's product. Fig. 12 displays the forecast for the last four years with the actual sales using the VAR model.

**Long Short-Term Memory Networks (LSTM):** Considering some of the learning from the models created for Taaka Vodka, a fewer number of models were tested for Jack Daniel's. The top-performing LSTM models were the models with price difference and base LSTM model without any auxiliary input. For the first mode, the price difference for the last available month was considered. All LSTM models were tested for 36 lags or lookback values, so a total of 72 models were created and tested for Jack Daniel's product. The model with price difference gave the minimum ASE of 3893.45 at a lookback value of 28, meaning that for predicting the next six months of sales, 28 previous sales were considered. Fig. 13 displays the forecast for the last four years with the actual sales using the LSTM model.

**Ensemble:** For Jack Daniel's ensemble model, a simple average ensemble approach was used. The difference in the results from VAR and LSTM models was significant. LSTM tends to over-fit with the small amount of overall training data; hence a simple ensemble in which both models contributed an equal amount to the final prediction seemed fit. The ensemble model attained an average ASE overall rolling windows of 6413.11, a 56% improvement over the Naive Model ASE. A graph of the ensemble forecast and the actual monthly cases sold of Jack Daniel's Whiskey is displayed in Fig. 14.

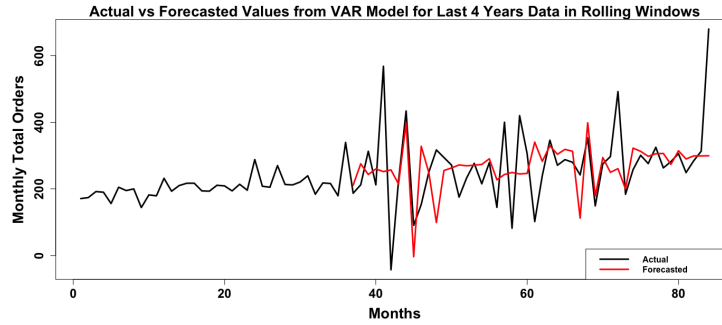


Fig. 12: VAR Model - Forecast of Jack Daniel's Total Cases per Month

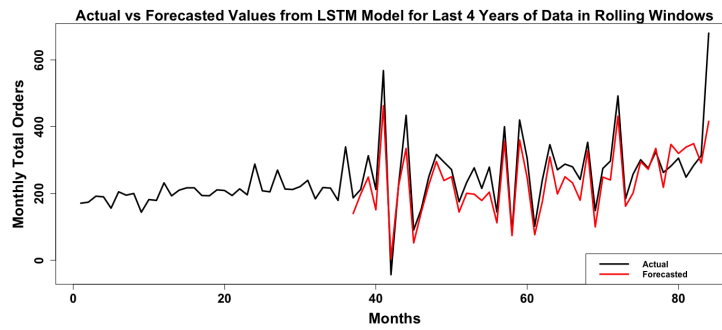


Fig. 13: LSTM Model - Forecast of Jack Daniel's Total Cases per Month

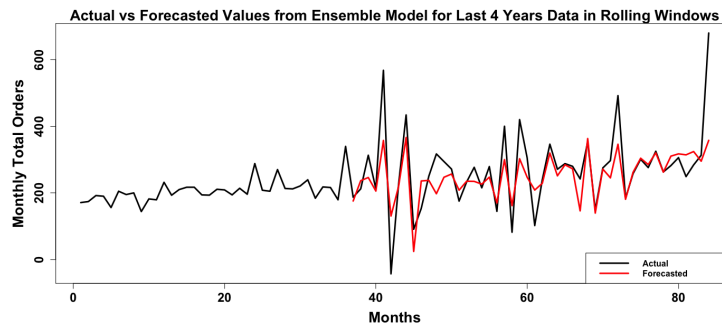


Fig. 14: Ensemble Model - Forecast of Jack Daniel's Total Cases per Months

Table 3: Summary of models tested and best ASE for Jack Daniel's Whiskey

Model Type	No. of Models Created and Tested	Best Identified Model ASE
Intuition based Naïve Model	1	14521.89
Univariate Model	N/A	N/A
VAR Model , Multivariate	52	14402.26
LSTM	72	3893.45
Ensemble	1	6413.11

Table 3 provides a summary of the total number of models created and tested for Jack Daniel's and the best-identified model with the smallest ASE for each of the model types.

## 8 Conclusion

In this study, analysis and forecasting was conducted using the last seven years of available monthly sales data for two alcoholic beverages, Taaka Vodka and Jack Daniel's Whiskey, each with a different time series. These products were the most sold and highest revenue grossing products, respectively, for the distribution center in question. Using ensemble models that harnessed naive, univariate, multivariate, and LSTM models, an overall reduction was achieved in the monthly forecasting error, as compared to the Naive Model for the total number of standard cases sold. Although LSTM models give minimum ASE, it tends to overfit, so ensemble model is preferred so that any overfitting or underfitting by one model gets compensated by other models. The monthly forecasting error, in terms of RMSE, was reduced by nearly 50% for Taaka Vodka and nearly 33.5% for Jack Daniel's Whiskey using the ensemble model. For the distributor, this improved forecasting translates to a real-world reduction of error in the number of cases sold each month. For Taaka Vodka, the ensemble model reduced monthly error from  $\pm 112$  cases to only  $\pm 60$  cases. For Jack Daniel's Whiskey, the ensemble model reduced monthly error from  $\pm 120$  cases to only  $\pm 80$  cases.

Additionally, it was found that models created for Taaka Vodka could not be generalized to the Jack Daniel's Whiskey. This led to the conclusion that there is no "one size fits all" model for all products sold by the distributor. Different products can have widely varying demand patterns and hence, command different models for sales forecasting. Understanding the individual sales data for the different products sold should lead to better efficiency in stocking, storage, and sales for Company A's distributor.

### 8.1 Future Work

Looking at the distributor perspective, they would be interested in the aggregate demand of the product for a month, that would help them maintain inventory without over-housing products in their warehouse. However, each client could

have different purchasing patterns, and hence, modeling demand at the individual retail customer level and not just at aggregate level could reveal different and interesting findings. This paper also lays the groundwork for the creation of an autoML framework that could create a slightly naive but useful model for each product and brand combination present in the dataset. Since these models would be generated in an automated fashion and not manually, these would provide significant value addition for the product house.

## References

1. Nick T. Thomopoulos. *Demand Forecasting for Inventory Control*. Springer, 2016.
2. Pamela Danese and Matteo Kalchschmidt. The role of the forecasting process in improving forecast accuracy and operational performance. *International Journal of Production Economics*, 131(1):204–214, 2011.
3. Michael Eisenstein. Hitting the mark. *Nature methods*, 11(9):894, Sep 2014.
4. Meritxell Venture-Cots, Ariel Watts, Monica Cruz-Lemini, Neil Shah, Nambi Ndugga, Peter McCann, A. S. Barritt, Anant Jain, Sanhita Ravi, Carlos Fernandez-Carrillo, Juan Abrales, Jose Altamirano, and Ramon Bataller. Colder weather and fewer sunlight hours increase alcohol consumption and alcoholic cirrhosis worldwide. *Hepatology*, 69(5), 16 October 2018.
5. Thomas F. Hogarty and Kenneth G. Elzinga. The demand for beer author(s): Thomas f. hogarty and kenneth g. elzinga source: The review of economics and statistics, vol. 54, no. 2 (may, 1972), pp. 195-198 published by: The mit press stable url: <https://www.jstor.org/stable/1926282> accessed: 14-11-2019 13:37 utc.
6. Zeynep H. Kilimci, A. O. Akyuz, Mitat Uysal, Selim Akyokus, M. O. Uysal, Berna Atak Bulbul, and Mehmet A. Ekmis. An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. *Complexity*, 2019:1–15, Mar 26, 2019.
7. Irem Islek and Sule G. Oguducu. A retail demand forecasting model based on data mining techniques. pages 55–60. IEEE, Jun 2015.
8. M.E. Nor, A.I.M. Nurul, and M.S. Rusiman. A hybrid approach on tourism demand forecasting. *Journal of Physics: Conference Series*, 995:12034, Apr 2018.
9. Dejan Mircetic, Svetlana Nikolicic, Marinko Maslaric, Nebojsa Ralevic, and Borna Debelic. Development of s-arima model for forecasting demand in a beverage supply chain. *Open Engineering*, 6(1), Nov 4, 2016.
10. Xinyu Hu, Paul Szerlip, Theofanis Karaletsos, and Rohit Singh. Applying svgd to bayesian neural networks for cyclical time-series prediction and inference. *arXiv.org*, 2019. ID: proquest2168396692.
11. Sangyeon Kim and Myungjoo Kang. Financial series prediction using attention lstm. *IDEAS Working Paper Series from RePEc*, 2019. ID: proquest2188166324.
12. G. P. Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.
13. Jamal Fattah, Latifa Ezzine, Zineb Aman, Haj El Moussami, and Abdeslam Lachhab. Forecasting of demand using arima model. *International Journal of Engineering Business Management*, 10:184797901880867, Oct 29, 2018.
14. Eric Zivot and Jiahui Wang. *Vector Autoregressive Models for Multivariate Time Series.*, pages 385–429. Modeling Financial Time Series with S-PLUS. Springer, New York, NY, 2006.

15. Henri Nyberg and Pentti Saikkonen. Forecasting with a noncausal var model. *Chemometrics and Intelligent Laboratory Systems*, 73(2):263, 2004.
16. Yixian Liu, Matthew Roberts, and Ramteen Sioshansi. A vector autoregression weather model for electricity supply and demand modeling. *Journal of Modern Power Systems and Clean Energy*, 6(4):763–776, Jul 2018.
17. Jason Brownlee. A gentle introduction to long short-term memory networks for the experts, May 2017. <https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts>.
18. Christopher Olah. Understanding lstm networks., August 2015. <https://colah.github.io/posts/2015-08-Understanding-LSTMs>.
19. Kenneth P. Burnham and David R. Anderson. Multimodel inference. *Sociological Methods Research*, 33(2):261–304, Nov 2004.
20. Anne B. Koehler and Rob J. Hyndman. *Another look at measures of forecast accuracy*, page 18. 2005.
21. Roy Yang. Omphalos, uber’s parallel and language-extensible time series backtesting tool, Jan 24 2018. <https://eng.uber.com/omphalos/>.
22. S. Aparna. Long short term memory and rolling window technique for modeling power demand prediction. pages 1675–1678. IEEE, Jun 2018.
23. Martha Starr-McCluer. The effects of weather on retail sales, January 2000.
24. Kyle B. Murray, Fabrizio Di Muro, Adam Finn, and Peter Popkowski Leszczyc. The effect of weather on consumer spending. *Journal of Retailing and Consumer Services*, 17(6):512–520, Nov 2010.
25. Climate data online - the national climatic data center. <https://www.ncdc.noaa.gov/cdo-web/>.