DEMAND FORECASTING FOR WHOLESALE SALES BY INDUSTRY CONSIDERING SEASONALITY IN DEMAND

by

Mohammad Wahidul Islam Murad, B.Sc. (Engr.) in Computer Science and Engineering,
Chittagong University of Engineering and Technology

A Major Research Project
presented to Ryerson University
in partial fulfillment of the requirements for the degree of

Master of Science
in the program of Data Science and Analytics

Toronto, Ontario, Canada, 2020

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Project. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Mohammad Wahidul Islam Murad

DEMAND FORECASTING FOR WHOLESALE SALES BY INDUSTRY CONSIDERING SEASONALITY IN DEMAND

Mohammad Wahidul Islam Murad

Master of Science 2020

Data Analysis and Analytics

Ryerson University

ABSTRACT

Demand forecasting is the basis for planning supply chain activities and is very important to choose effective forecasting technique that is appropriate on specific data set. The appropriate forecasting technique helps management to use this information and maintain the flow of materials, products and information in supply chain management. Many active researches are going on different demand forecasting techniques for several years. The aim of this research project is to study and implement effective forecasting techniques applied on time-series data set with different wholesale products by industry type under the North American Industry Classification System. The objective of this research project is to demand forecast for wholesale product by industry based on historical time-series data, evaluate and compare forecast accuracy using performance measurement evaluation metrics. In this research project, three time-series forecasting models ARIMA, SARIMAX and Seasonal Decomposition were used to predict the demand for 23 different wholesale products. The evaluation metrics Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) were used to identify the accuracies between actual and predicted data. The outcome of this project is to compare different forecasting models and identify the most suitable forecasting technique that can be used for predicting wholesale products.

Keywords: North American Industry Classification System, Supply chain management, ARIMA, SARIMAX, Seasonal Decomposition, RMSE, MAPE.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude and due respect to my supervisor **Dr. Saman Hassanzadeh Amin** for successful completion of Major Research Project, a Milestone of Master of Science in Data Science and Analytics program.

I am very grateful to him for his valuable suggestions and continuous supports which helped me a lot during the period of my project work. He has been a great support throughout the term to guide and direct my research and provide valuable feedback. I would like to appreciate him for always providing important suggestions.

Thank you, Professor Dr. Saman Hassanzadeh Amin.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES AND CHARTS

1. Introduction

Supply chain and demand forecasting have been developing as a result of economic globalization, technology development and growing consumer demand. Among various supply chain processes, supply chain forecast and measuring its accuracy has been one of the major focus areas drawing attention from both researchers and business holders. This MRP is intended to predict demand for different wholesale products by industry type under the North American Industry Classification System. The demand forecast is based on historical time-series data and perform financial estimation in supply-chain management system (Wang and Chen, 2019). The raw data on sales and inventory over time has significant role for the implementation of multi-product demand forecasts (Boone et al. 2019). In addition to demand forecast compare performances based on different forecasting model and machine learning algorithms.

1.1. Background

Supply chain performance is the important part for many companies which is directly related to the company revenues and business performance. Forecasting plays an important role to measure uncertainty and risk associated to the businesses. Recent research has shown that forecasting enables improvements in supply chain performance and develop present and future strategy of the business (Boone et al. 2019). The role of demand forecasting in supply chain is the key to provide the right direction. Many important progresses related to demand processes, demand inference and demand forecasts have been made (Syntetos et al. 2016). In this Major Research Project, financial estimation in supply chain management system is performed and compared performances based on different forecasting model and machine learning algorithms.

1.2. Research Question

The questionnaire of this research project is divided into three parts. The first part contains descriptive study about supply chain system and various demand forecasting model. The second part of the study is to classify training and test classes that could enable predictive analytics and to implement different forecasting model and machine learning algorithm. The third part of the study is to compares performances among different evaluation metric and to identify the contributing factors related to the performances. The dataset that already collected from Statistics Canada and hope it will help to get the consistent result and insight into the real business performance analysis.

1.3. Independent/Dependent Variables

The list of all the available fields in the dataset is given in Appendix C.

2. LITERATURE REVIEW

The main idea of this project is to demand forecasting for wholesale sales by industry and comparing prediction performances with the use of different machine learning algorithms. Although there have been several works done in the area of demand forecasting in supply-chain management system, this prediction for wholesale sales by industry based on historical time-series data is an innovative idea which is implemented in this paper and compared performances using different machine learning algorithms.

Demand forecasting refers to the process of predicting the future demand for the product which are categorized as different types of industry where products were produced or categorized as industry type under the North American Industry Classification System (NAICS). In this section, an overview of articles and journals as a reference for this project is provided.

The demand forecasting and financial estimation in supply chain management system are significant in industries and businesses. There are several studies done on demand forecasting and financial estimation in supply chain management system; some of which are related to this project. One of the most related articles published by Wang and Chen (2019) proposed a demand-pull framework to perform demand forecast and financial estimation considered the interactive dynamics of semiconductor companies. They proposed ARIMA (Autoregressive Integrated Moving Average) model for demand forecasting, ARIMA and VAR (Vector Auto-Regression) for financial estimation. Finally, they have compared different demand forecasting technique and compared performances. They have also studied time-series models which is very helpful to get insight the historical data and to predict the future values of the outcome. Roque et al. (2019) mentioned that ARIMA model is a simplification of an ARMA and SARIMA (Seasonal Autoregressive Integrated Moving Average) model where non-stationarity time-series and seasonality are considered. In this research project ARIMA is one of the models used for demand forecasting of wholesale sales data. Syntetos et al. (2016) studied about supply chain forecasting and mentioned the methodology of different supply chain and demand forecasting techniques.

They have suggested ARIMA model for upstream propagation of demand. Also, they have suggested ARIMA and INARMA (Integer Auto-Regressive Moving Average) for supply chain forecasting.

Lin et al. (2019) introduced statistic-based model Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX) and according MSE results they found that SARIMAX have similar average MSE and it performs reliably better than Support Vector Regressor (SVR) and Linear Regression. Van et al. (2020) mentioned that the more effective forecasting techniques are SARIMAX, SARIMA, Holt-Winters and Seasonal Decomposition model (DM).

Bruzda (2019) re-examined the exponential smoothing model in the area of supply chain and logistics forecasting with microeconomic time-series data and proposed that quantile SES-d (Simple Exponential Smoothing with Drift), QARMA (Quantile ARMA), Asymmetric SES-d and SES-d with smoothing the MAD (Mean Absolute Deviation) and an optimized δ models can be used in demand forecasting. The parameters of these models were optimized considered to the LINLIN loss function and applied to the quantile regression estimator. Pereira et al. (2018) propose a forecasting model to identify the variables that impact the return of scrap tires based on correlation between the number of tires added in the market and number of tires directed to the destination.

Rivera and Burnaev (2017) examined the Guassian Processes to evaluate performances of commercial data from point of sales having large size and high dimensionality. The study suggested that Guassian Processes modeling is very helpful for fast moving consumer industry as a policymaking tool for management. Van et al. (2020) introduced profit function that combined forecasting accuracy with business proficiency. They have studied the processes of high-level strategic sales forecast by compared 35 times series datasets with 17 different forecasting techniques. Tanizaki et al. (2019) proposed store-specific demand forecasting model using different machine learning algorithm (e.g. Bayesian Linear Regression, Boosted Decision Tree Regression and Decision Forest Regression) where store location, the weather and events were considered.

In the case of Neural Network approach, historical data can be analyzed considering seasonality. Kechyn et al. (2018) participated in sales forecasting competition and examined the solution of the machine learning problem based on time-series data. Finally, they have proposed a highly effective method of using CNN (Convolutional Neural Network) WaveNet model to perform sales forecasting. Chen et al. (2019) studied the retailer sales forecasting on Tmall and designed two mechanisms for sales forecasting which are seasonality extraction and distribution transformation. They have applied these methods to classic regression models in neural network and Gradient Boosting Decision Tree. Karb et al. (2020) proposed an analytic approach to investigate the efficiency of network-based Transfer Learning under deep neural networks for food retail sales forecasting. In addition, Hirt et al. (2020) studied about similar Transfer Learning and proposed a generalized model based on additive regression models that introduced new entities that better than existing models. Lee et al. (2019) proposed a baseline model Temporal-Guided Network (TGNet) with graph networks and temporal-guided embeddings for short-term demand forecasting. Zhao and Wang (2017) proposed an integrated automated CNN (Convolutional Neural Network) framework to forecast sales based on E-commerce raw log data. It gives details idea about sales forecast in E-commerce where CNN is used to predict sales from online shopping. Moreover, Bandara et al. (2019) proposed a systematic pre-processing framework using a Long Short-Term Memory (LSTM) neural network and introduce a product grouping strategy to enhance the LSTM learning schema. Abbasimehr et al. (2020) proposed a demand forecasting method based on multi-layer LSTM networks that automatically chooses the best forecasting model considered different combinations of LSTM hyperparameters. They had evaluated the performances based on RMSE and Symmetric Mean Absolute Percentage Error (SMAPE).

Rivera-Castro et al. (2019) proposed demand forecasting techniques for build-to-order supply chains with data transformation technique and the results from thirteen forecasting methods proved that researchers can work to this area. Abolghasemi et al. (2020) investigated 843 real demand time series with different values of coefficient of variations (CoV) and proposed a hybrid model with ARIMA with covariate (ARIMAX). They indicate that this statistical and machine learning model can be used to forecast volatile demand series data.

Lin et al. (2019) studied sentiment-sales and related data and introduced a pre-trained BERT (Bidirectional Encoder Representations from Transformers) language model architecture to

perform sales prediction. In the case of evaluation metric Mean Squared Error (MSE) was used to compare different forecasting model. Van Belle et al. (2020) investigated different modeling approaches, forecasting methods and compared different method to improve short-term forecast accuracy by anticipating sell-through data into the forecasting process. Jiao et al. (2020) investigated tourism forecasting accuracy and proposed a spatiotemporal autoregressive model incorporated both spatial dependence and spatial heterogeneity to improve forecasting accuracy based on MAPE and Mean Absolute Scaled Error (MASE).

In the case of measuring the forecast accuracy Root Mean Squared Error (RMSE) is mostly used metrics to evaluate forecast models. Wang and Chen (2019), Roque et al. (2019) utilized this method to evaluate forecast models. Xu and Sharma (2017) indicated that multiple regressions, time-series analysis, random forest and boosting tree were executed in parallel due to recent advances in computation power, software development and capability of executing in parallel. They have also proposed an ensemble method nominated from smaller validations errors and applied MAPE (Mean Absolute Percentage Error) to classify error. Wang and Chen (2019) also used MAPE to assess the forecasting errors. Joanna Bruzda (2019) mentioned that MAPE is the best to evaluate forecasts performances for monthly time-series data. Roque et al. (2019) mentioned that Mean Absolute Error (MASE) is better than the average one-step naive forecast method. Calster et al. (2020) defined the RMSE, MAPE and MASE equations and performed evaluation measures for 26 different forecast models. Hirt et al. (2020) evaluated forecast error for Transfer Learning algorithm using two mostly used metrics RMSE and MAPE.

Rivera and Burnaev (2017) applied RMSLE (Root Mean Squared Log Error) to evaluate the performance of Guassian Processes. Kechyn et al. (2018) applied NWRMSLE (Normalized Weighted Root Mean Squared Logarithmic Error) as evaluation metric for CNN WaveNet architecture. Karb et al. (2020) used MSE (Mean Squared Error) as evaluation metric for Transfer Learning network model.

Boone et al. (2019) reviewed the impact of big data on product forecasting and mentioned different sources of big data, forecasting methodology and organizational challenges which is very helpful for this research project to demonstrate the idea of identifying the potential sources of dataset and designed the model to compare the different forecasting techniques.

Before comparing the performances of demand forecasting of wholesale products, it is important to explore the dataset and learn about structure, format and information about the data. A detailed exploratory analysis of the dataset was conducted using reference from visual representation analysis and technique. After explored the dataset, the next step is to compare the performances of demand forecasting using different machine learning algorithm.

3. PROBLEM STATEMENT

Today's competitive global market rely on effective business planning and materials flow along with the supply chain management system. Many industries have put many efforts my making significant improvements by using a suitable method that supports and facilitate the process of supply chain management. One of the important aspects of supply chain management system is to enable an effective forecasting technique that helped to produce accurate, timely estimates of future product demand.

In this research project the focus is to forecast demand for different wholesale product by industry type under the North American Industry Classification System (NAICS). To identify the effective forecasting techniques and measured error between actual and predicted result. The dataset was collected from Statistics Canada website. The dataset This dataset is monthly wholesale trade survey data and it is open-sourced. In fact, improved demand forecasting accuracy can help the industry in financial savings, enhanced competitiveness, improved channel relationships and customer satisfaction.

4. EXPLORATORY DATA ANALYSIS

The dataset that is selected for this project contains wholesale sales for different industries based on North American Industry Classification System (NAICS). There are some rows that don't have any data and as it will affect to make the pivot table in python, filled null value with zero. The data were collected based on Monthly Wholesale Trade Survey and the survey code is 2401. The dataset has three dimensions as below.

1. Geography
2. Sales, price and volume
3. North American Industry Classification System (NAICS).

4.1. Data Acquisition

Data for this Major Research Project collected from Statistics Canada (www150.statcan.gc.ca) website. The datasets are open-sourced and compliant with the MRP requirements.

4.2. Data Source and Data Files

The datasets have been downloaded from Statistics Canada website. For predictive analytics, the dataset will be split into two subsets for training and test to evaluate the result. The details about training/testing of datasets are given below.

- As dataset contains time-series data for historical dates, I have selected year from 2009 to 2015 as training set.
- The data from 2016 to 2019 selected as test set.

Data files are posted on GitHub. The link to the GitHub repository and a sample file, both are included in Appendix B.

4.3. Data Structure Analysis

The main part of the data is the industry type and value of monthly transactions. The industries were classified based on North American Industry Classification System (NAICS). The sectors of the industries are defined based on definition of Statistics Canada.

The sector contains establishments primary engaged in wholesaling merchandise and rendering services incidental to the sale of merchandise. The wholesaling process is an intermediate step in the distribution of goods. The details about sector structure are given below.

Sector Structure: The subsector has 3-digit level code and industry groups are defined with 4-digit level code under each subsector.

Table 1 represents the wholesale sale industry sector structure and classifications based on North American Industry Classification System (NAICS).

Table 1 – Wholesale sale industry sector structure

[41] Wholesale trade		
[411] Farm product merchant wholesalers		
[413] Food, beverage and tobacco merchant wholesalers		
	[4131] Food merchant wholesalers	
	[4132] Beverage merchant wholesalers	
	[4133] Cigarette and tobacco product merchant wholesalers	
[414] Personal and household goods merchant wholesalers		
	[4141] Textile, clothing and footwear merchant wholesalers	
	[4142] Home entertainment equipment and household appliance merchant wholesalers	
	[4143] Home furnishings merchant wholesalers	
	[4144] Personal goods merchant wholesalers	
	[4145] Pharmaceuticals, toiletries, cosmetics and sundries merchant wholesalers	
		[41451] Pharmaceuticals and pharmacy supplies merchant wholesalers
		[41452] Toiletries, cosmetics and sundries merchant wholesalers
[415] Motor vehicle and motor vehicle parts and accessories merchant wholesalers		
	[4151] Motor vehicle merchant wholesalers	
	[4152] New motor vehicle parts and accessories merchant wholesalers	
	[4153] Used motor vehicle parts and accessories merchant wholesalers	
[416] Building material and supplies merchant wholesalers		
	[4161] Electrical, plumbing, heating and air-conditioning equipment and supplies merchant wholesalers	
	[4162] Metal service centres	
	[4163] Lumber, millwork, hardware and other building supplies merchant wholesalers	
[417] Machinery, equipment and supplies merchant wholesalers		
	[4171] Farm, lawn and garden machinery and equipment merchant wholesalers	
	[4172] Construction, forestry, mining, and industrial machinery, equipment and supplies merchant wholesalers	
	[4173] Computer and communications equipment and supplies merchant wholesalers	
	[4179] Other machinery, equipment and supplies merchant wholesalers	
[418] Miscellaneous merchant wholesalers		
	[4181] Recyclable material merchant wholesalers	
	[4182] Paper, paper product and disposable plastic product merchant wholesalers	
	[4183] Agricultural supplies merchant wholesalers	
	[4184] Chemical (except agricultural) and allied product merchant wholesalers	
	[4189] Other miscellaneous merchant wholesalers	

4.4. Data Analysis

The dataset column North American Industry Classification System (NAICS) was explored and in this column, there are 7 subsector and 23 industry groups data.

Figure 1 represents the yearly total wholesale sales for all wholesale sales product from year 2009 to 2019. According to the trend it is clearly visible wholesale sales gradually increased over the trend and there is a sharp peak at year 2017.

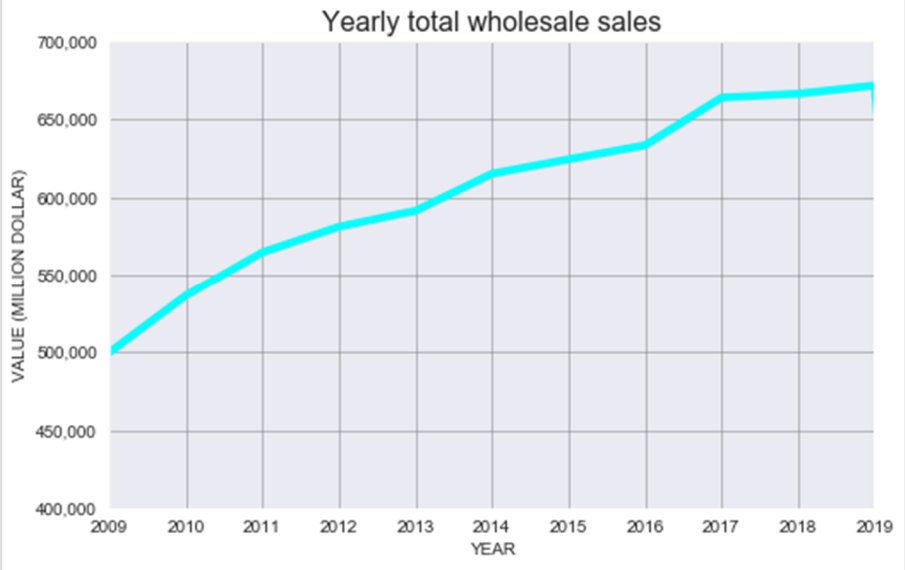


Figure 1 – Yearly total wholesale sales

Figure 2 depicts the percentage of industrial subsector of total wholesale sales and it shows that the highest percentage is 20.4% for [417] Machinery, equipment and supplies merchant wholesalers.

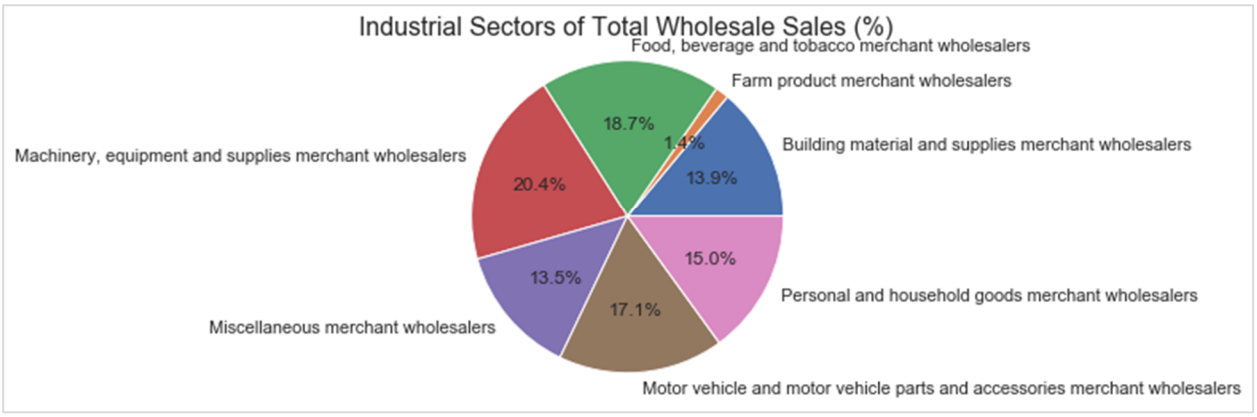


Figure 2 – Percentage of industrial subsector of total wholesale sales

Figure 3 below depicts the average wholesale values (Million Dollar) by subsector. Looking at the chart the sector subsector related to [417] Machinery, equipment and supplies merchant wholesalers has the highest number of values averaged over the trend.

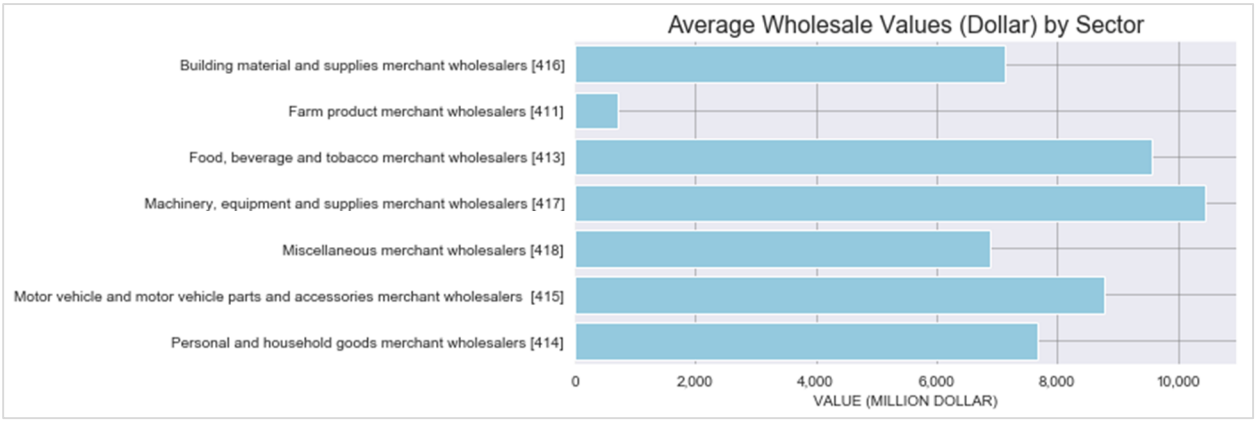


Figure 3 – Yearly sectors trend with price

Figure 4 below depicts the average wholesale values (Million Dollar) by industry group. Looking at the chart the industry group related to [4131] Food merchant wholesalers industry has the highest number of values averaged over the trend.

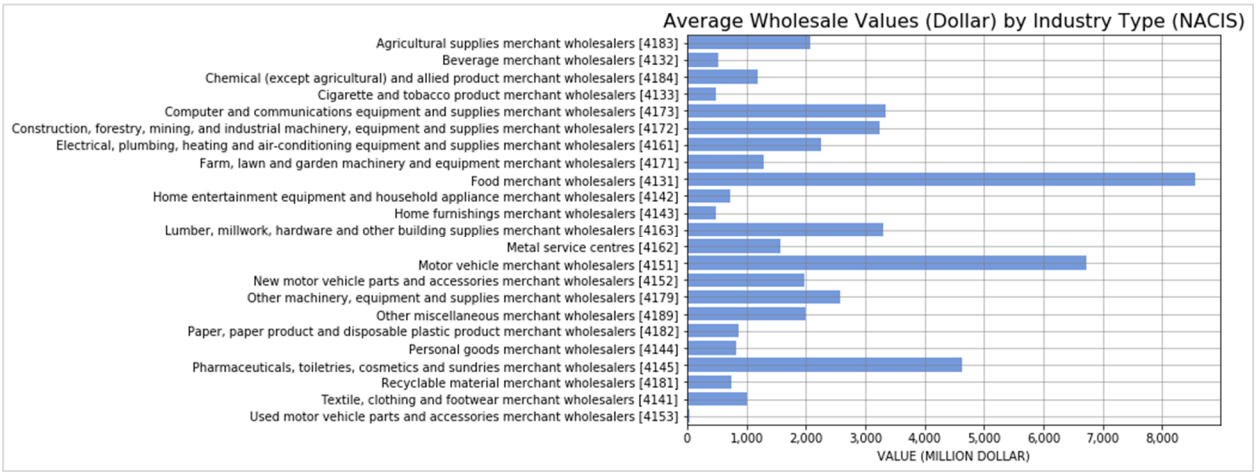


Figure 4 – Average wholesale sales price by industry type

Figure 5 below illustrates the boxplot representation to highlight minimum and maximum values for wholesales of each industry group. The industry group related to [4131] Food merchant wholesalers industry has the highest number of outliers.

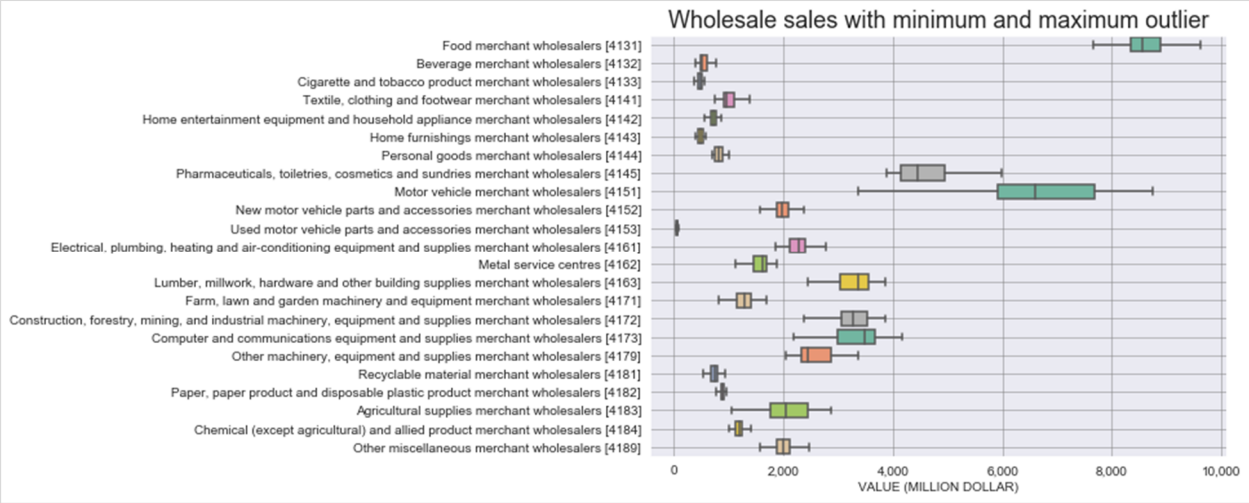


Figure 5 – Boxplot to represent industries wholesale sales with minimum and maximum outlier

4.5. Extended Exploratory Analysis

In the detail exploratory analysis, the pivot table designed for each industrial group and depicts the value for each year (from 2009 to 2019).

Table 2 represents the pivot table that helps to understand the wholesale sale values by industry group and by each year.

Table 2 – Wholesale sales values (Million Dollar) by industry group and by each year

North American Industry Classification System (NAICS) YEAR	Agricultural supplies merchant wholesalers [4183]	Beverage merchant wholesalers [4132]	Chemical (except agricultural) and allied product merchant wholesalers [4184]	Cigarette and tobacco product merchant wholesalers [4133]	Computer and communications equipment and supplies merchant wholesalers [4173]	Construction, forestry, mining, and industrial machinery, equipment and supplies merchant wholesalers [4172]	Electrical, plumbing, heating and air-conditioning equipment and supplies merchant wholesalers [4161]	Farm, lawn and garden machinery and equipment merchant wholesalers [4171]	Food merchant wholesalers [4131]	Home entertainment equipment and household appliance merchant wholesalers [4142]
2009	$15,416.00	$5,064.00	$12,814.00	$6,203.00	$28,340.00	$29,382.00	$23,028.00	$12,064.00	$93,566.00	$7,373.00
2010	$20,211.00	$5,217.00	$13,599.00	$6,381.00	$31,876.00	$34,751.00	$24,321.00	$11,015.00	$97,415.00	$8,113.00
2011	$22,620.00	$5,692.00	$14,212.00	$6,189.00	$34,788.00	$42,742.00	$25,058.00	$13,648.00	$99,655.00	$8,644.00
2012	$20,635.00	$5,802.00	$14,435.00	$6,106.00	$38,168.00	$44,158.00	$25,794.00	$14,708.00	$101,907.00	$8,602.00
2013	$22,152.00	$5,846.00	$13,072.00	$5,935.00	$41,705.00	$42,290.00	$26,770.00	$15,770.00	$103,057.00	$8,902.00
2014	$23,524.00	$5,926.00	$13,581.00	$5,533.00	$43,735.00	$43,727.00	$29,072.00	$15,920.00	$102,785.00	$9,142.00
2015	$27,401.00	$6,336.00	$14,545.00	$5,018.00	$44,240.00	$39,548.00	$28,290.00	$15,291.00	$101,609.00	$9,294.00
2016	$29,880.00	$6,841.00	$15,769.00	$5,074.00	$43,752.00	$35,883.00	$27,935.00	$15,899.00	$105,785.00	$8,065.00
2017	$30,108.00	$7,392.00	$16,156.00	$5,356.00	$41,477.00	$38,182.00	$28,980.00	$19,237.00	$106,420.00	$8,952.00
2018	$29,682.00	$7,707.00	$14,540.00	$5,758.00	$42,262.00	$38,971.00	$29,344.00	$18,895.00	$108,027.00	$9,442.00
2019	$29,753.00	$7,809.00	$14,651.00	$5,517.00	$48,152.00	$39,402.00	$29,406.00	$16,965.00	$107,637.00	$9,252.00

New motor vehicle parts and accessories merchant wholesalers [4152]	Other machinery, equipment and supplies merchant wholesalers [4179]	Other miscellaneous merchant wholesalers [4189]	Paper, paper product and disposable plastic product merchant wholesalers [4182]	Personal goods merchant wholesalers [4144]	Pharmaceuticals, toiletries, cosmetics and sundries merchant wholesalers [4145]	Recyclable material merchant wholesalers [4181]	Textile, clothing and footwear merchant wholesalers [4141]	Used motor vehicle parts and accessories merchant wholesalers [4153]	All
$19,126.00	$27,545.00	$23,940.00	$8,370.00	$10,460.00	$48,944.00	$8,759.00	$9,482.00	$0.00	$499,860.00
$20,546.00	$25,891.00	$24,091.00	$10,472.00	$9,863.00	$49,012.00	$8,608.00	$10,487.00	$585.00	$536,520.00
$22,741.00	$27,607.00	$24,502.00	$10,799.00	$9,039.00	$49,970.00	$9,426.00	$11,135.00	$591.00	$564,627.00
$23,863.00	$29,035.00	$22,903.00	$10,756.00	$8,676.00	$49,553.00	$8,561.00	$11,073.00	$558.00	$581,101.00
$25,062.00	$29,811.00	$21,282.00	$10,437.00	$8,532.00	$50,698.00	$8,177.00	$11,884.00	$696.00	$591,350.00
$27,986.00	$28,882.00	$21,600.00	$10,753.00	$9,108.00	$53,149.00	$10,698.00	$12,231.00	$716.00	$615,162.00
$27,490.00	$29,219.00	$22,681.00	$11,129.00	$9,787.00	$56,209.00	$10,861.00	$11,897.00	$671.00	$624,407.00
$23,566.00	$29,684.00	$24,130.00	$10,595.00	$9,959.00	$57,977.00	$7,087.00	$11,399.00	$600.00	$633,419.00
$23,554.00	$34,700.00	$25,616.00	$11,115.00	$10,988.00	$59,794.00	$8,525.00	$13,923.00	$687.00	$664,050.00
$23,915.00	$38,732.00	$25,362.00	$10,988.00	$11,067.00	$64,003.00	$9,356.00	$14,266.00	$781.00	$666,548.00
$24,343.00	$38,483.00	$27,496.00	$10,230.00	$10,925.00	$68,080.00	$8,513.00	$14,557.00	$830.00	$671,841.00

5. METHODOLOGY AND EXPERIMENTS

5.1. Aim of Study

In this research project, different demand forecast model implemented to forecast demand for wholesale sales data based on industry type of North American Industry Classification System (NAICS). Firstly, different forecasting model to predict the demand of wholesale sales product and compared the performances based on error produced from actual and predicted data. To select the best forecasting model, 3 different models were used with different parameter settings. Finally, the goal is to evaluate the performances of each model by using 2 different evaluation metrics.

5.2. Dependent and Independent Variable(s)

In this experiment, the dependent variables are REF_DATE, GEO, SCALAR_FACTOR, COORDINATE, Sales price and volume, North American Industry Classification System (NAICS) and VALUE. The North American Industry Classification System (NAICS) column further divided based on industry subsector and industry group. There are other fields in the dataset and following are used as inputs to train, validate and test the forecast models.

- REF_DATE: The month and the year information.
- North American Industry Classification System (NAICS): Industry subsector and industry group information.
- VALUE: Transaction value in million dollars.

5.3. Factors and Levels

In this experiment, the factors are the different statistical models with different parameter settings that can validate the model algorithm and its evaluation metric performance.

5.4. Experimental Design

a. Data Preprocessing

The dataset from Statistics Canada called “Monthly Wholesale Trade Survey” is monthly survey data, it has three dimensions which are “Geography”, “Sales, price and volume” and “North American Industry Classification System (NAICS)”. The first step is to copy the data from csv file to a Python data frame and fill missing value with zero. Also, it is required to convert the values into float type to facilitate for further calculation. The next step is to divide the dataset based on industry subsector and industry group.

b. Feature Selection and Experimental Design

The dataset was divided into three parts, one for train set from year 2009 to 2017, validation set from year 2016 to 2017 and test set from year 2018 to 2019. The experimental design architecture is shown in figure 6.

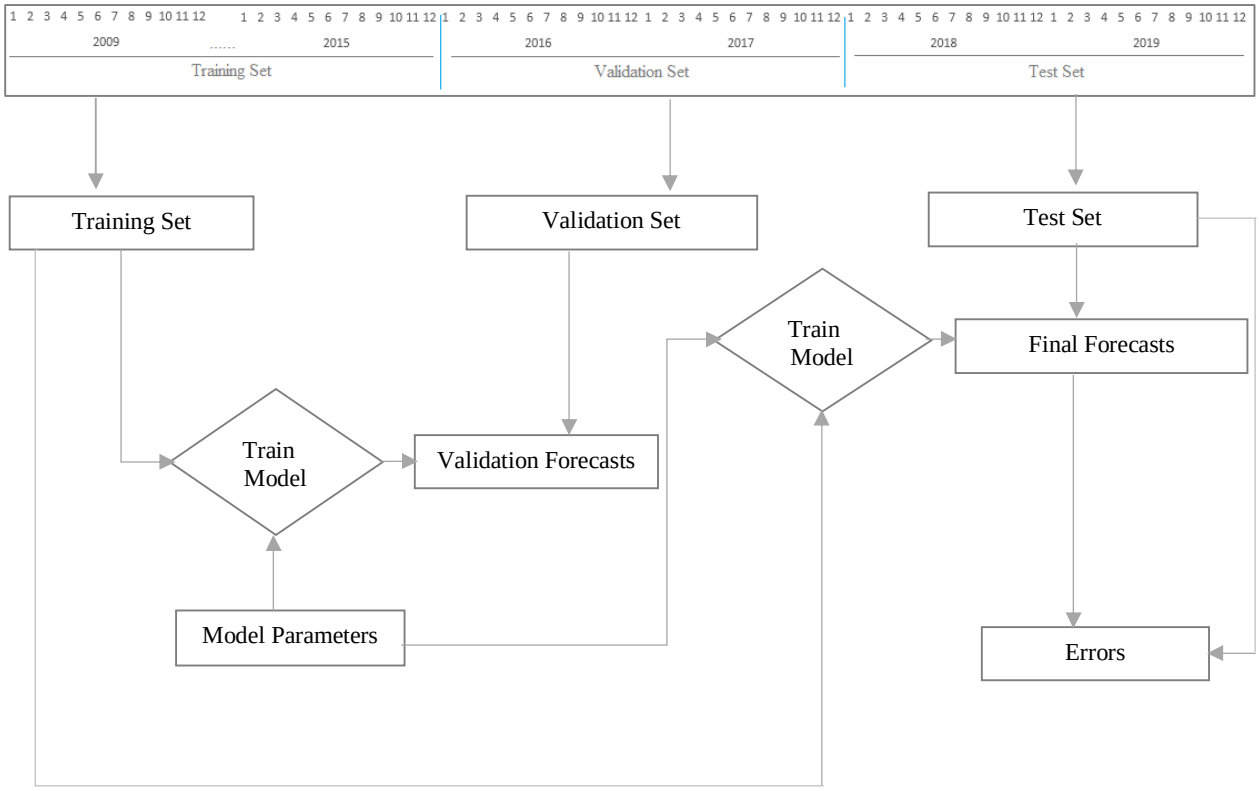


Figure 6 – Experimental design

c. Cross Validation

It is required to implement a validation technique as it is required to compare different forecasting model. In this project, time series cross-validation technique used to measure the accuracy using RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) between actual and predicted data. Finally, the best forecasting model will be selected with the smallest RMSE and MAPE values.

5.5. Experiment Performance and Revisions

Different forecasting model were analyzed and implemented best forecasting model that can help to fine tune the model, improve model accuracy and produce smallest number of errors. Details of different demand forecasting model implementation are given below.

a. Experiment 1

In the first experiment of this project, ARIMA (Autoregressive Integrated Moving Average) model for demand forecasting is used. An ARIMA model is a class of statistical models for analyzing and forecasting time series data. ARIMA model deals with univariate data and consider each time series separately and predict them in isolation.

Wang and Chen (2019) mentioned that, there are two kinds of ARIMA model static and dynamic. Static ARIMA forecast consider results of previous forecast as a predictor of future inflation outcomes. On the other hand, dynamic ARIMA synchronously considers the historical values of the predictors and the outcome. Kazemzadeh et al. (2020) indicated two functions Autocorrelation (ACF) and Partial Autocorrelation (PACF) which helps to determine if the time series is stationary or dynamic. ARIMA model uses the dependant relationship between an observation and lagged observations and integrated the observations for present and previous time step. Also, it considered the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The standard notation of ARIMA model is ARIMA (p, d, q) where the parameters are referenced with integer values to indicate the specific ARIMA model. Pereira et al. (2018) described the ARIMA model parameter, where the value of p is the number of autoregressive (AR) terms, d is the number of differences for raw observations and q is the size of moving average (MA) window.

In this experiment, the target is to fine tune the p, d and q parameter to get the better forecast result and to get minimum prediction error between actual and predicted data. Based on the accuracy evaluation, final ARIMA parameter will determine and implement the prediction model between train and test sets of data set.

b. Experiment 2

In the second experiment, SARIMAX (Seasonal Auto Regressive Integrated Moving Averages with eXogenous regressor) model is used to predict the demand for wholesales data which is one of the best performing models. Lin et al. (2019) mentioned that, SARIMAX model is an extension of SARIMA model with external variables. This model denotes by SARIMAX (p, d, q) (P, D, Q) s(X), where p, d, q are orders of autoregressive, difference and moving average. Moreover P, D, Q are orders of seasonal autoregressive, difference and moving average. X is the external variable and S is the seasonal period. The main difference to the ARIMA model is that it integrates external

variable to the model which is help to preferable in a business environment where different objectives need to establish for predicting time series data.

Van et al. (2020) represents the overview of different forecasting techniques and mentioned that SARIMAX consider the seasonality of time series data and AR (Autoregressive), MA (Moving average), SAR (Seasonal Autoregressive) and SMA (Seasonal Moving Average) are the important hyper parameters that helps to determine the final parameter settings of SARIMAX model.

In this experiment, the goal is to fine tune the p, d and q parameter to get the better forecast result and to get minimum prediction error between actual and predicted data. Based on the accuracy evaluation, final SARIMAX parameter will determine and implement the prediction model between train and test sets of data set.

c. Experiment 3

In the case of third experiment, Seasonal Decomposition model used for demand forecasting. Rivera and Burnaev (2017) examined the daily log demand time series for Fortune 500 company data set and informed about the possibility to observe a seasonality component of the time series data. The Seasonal Decomposition model automatically decompose time series data in Python and breaks into systematic and unsystematic components. The component of the Seasonal Decomposition model is (O, T, S, E). The value of O refers to the observed or original value of the trend, T is the trend, S is the seasonality and E refers to the residual errors.

The seasonal decomposition can be modeled either additive or multiplicative. Abolghasemi et al. (2020) mentioned that a time series is defined by the model Y_t, the additive form of the model is $Y_t = T_t + S_t + E_t$ and multiplicative form of the model is $Y_t = T_t * S_t * E_t$where T_t represents the trend component, S_t represents the seasonal component and E_t represents the errors.

Van et al. (2020) examined that the training of Seasonal Decomposition takes the least amount of time which is 0.05 seconds.

5.6. Measuring Prediction Model Performance

To get the evaluation metric, Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are the primary metrics to interpret the forecasting models. RMSE is the standard deviation of the prediction errors of different models. MAPE is a statistical measure to calculate the accuracy as a percentage to evaluate the forecasting models. MAPE is the difference between

actual and forecast value divided by the actual value. Calster et al. (2020) defined RMSE and MAPE equations and performed evaluation measures for 26 different forecast models.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(F_t - A_t)^2} \quad (1)$$

In equation (1) of RMSE, A_t represents actual value, F_t represents forecast value and n represents number of samples.

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| * 100 \quad (2)$$

In equation (2) of MAPE (%), A_t represents actual value, F_t represents forecast value and n represents number of samples.

5.7. Forecasting Models and Comparison

In this project, statistical models with machine algorithms used to forecast the time-series wholesale sales data. The model with minimum number of errors between actual and forecast values will be select to predict the wholesale sales data and as the final suggested model to use. Figure 7 represents the high-level diagram for forecasting techniques and error measurements.

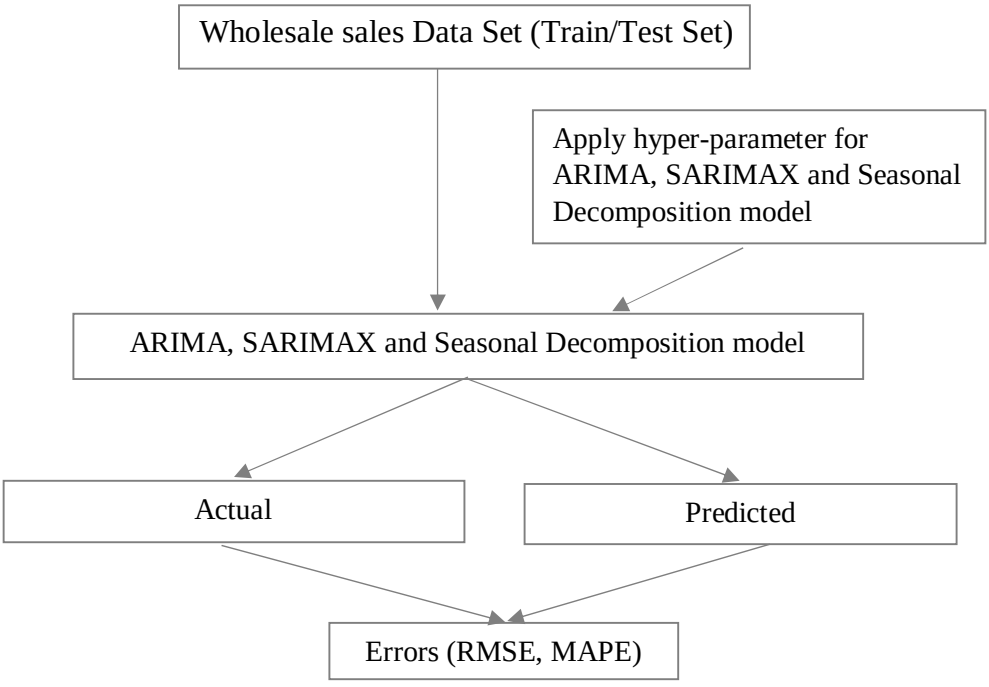


Figure 7 – Forecasting techniques and error measurements

6. RESULTS AND DISCUSSION

6.1. Data Exploratory Analysis Results

The exploratory analysis of this project helps to categorize the data according to different industry type and pivot table to sum up the values. The different industry has different values and fluctuations over the trend and exploratory result helps to understand about the feature of the dataset and major fluctuations. The prediction models were selected based on wholesales sales values and time series trend characteristics which helped to get the minimum number of errors between actual and predicted values.

6.2. Cross Validation Results

In this project, the data from year 2016 and 2017 were selected as validation set and measured the accuracy using RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) for actual and predicted data. The prediction was performed based on different parameter settings of ARIMA, SARIMAX and Seasonal Decomposition and identified the parameter that provide minimum RMSE and MAPE (%) values. Similarly, the prediction models were applied with selected parameter settings and on train and test data set.

According to the different combination of p, d and q parameter for ARIMA model, the parameter settings (0, 1, 0) better suit for this data set since the series is not stationary and it provides minimum RMSE and MAPE. In the case of SARIMAX model, the (p, d, q) as (6, 1, 1) gives minimum RMSE and MAPE. Moreover, for Seasonal Decomposition model the parameter 'freq' selected as 12 because dataset contains monthly data.

Table 3 represents the measurement of RMSE and MAPE (%) for ARIMA, SARIMAX and Seasonal Decomposition model applied on train and validation set of wholesale sale data.

Table 3 – RMSE and MAPE for ARIMA, SARIMAX and Seasonal Decomposition models

North American Industry Classification System (NAICS)	ARIMA_RMSE	ARIMA_MAPE	SARIMAX_RMSE	SARIMAX_MAPE	Seasonal Decompose_RMSE	Seasonal Decompose_MAPE
Total Wholesale Sales	604.863133	2.966778	604.373967	0.818043	682.808871	0.735513

Figure 8 represents the ARIMA model prediction for total wholesale sales values applied on train and validation set of data set.

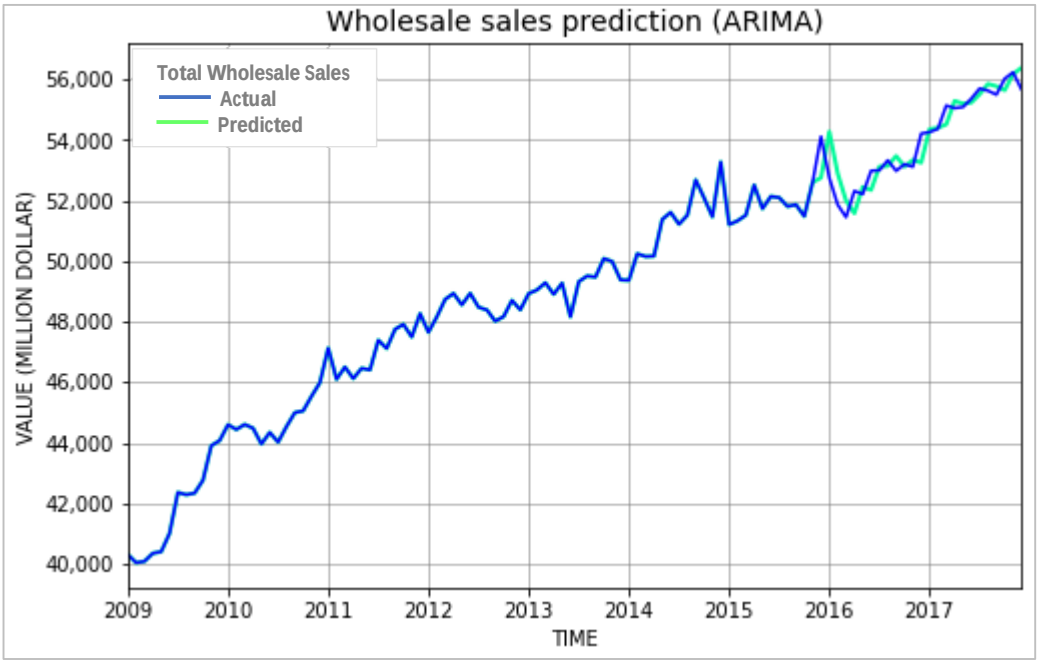


Figure 8 – ARIMA model prediction for total wholesale sales values

Figure 9 represents the SARIMAX model prediction for total wholesale sales values applied on train and validation set of data set.

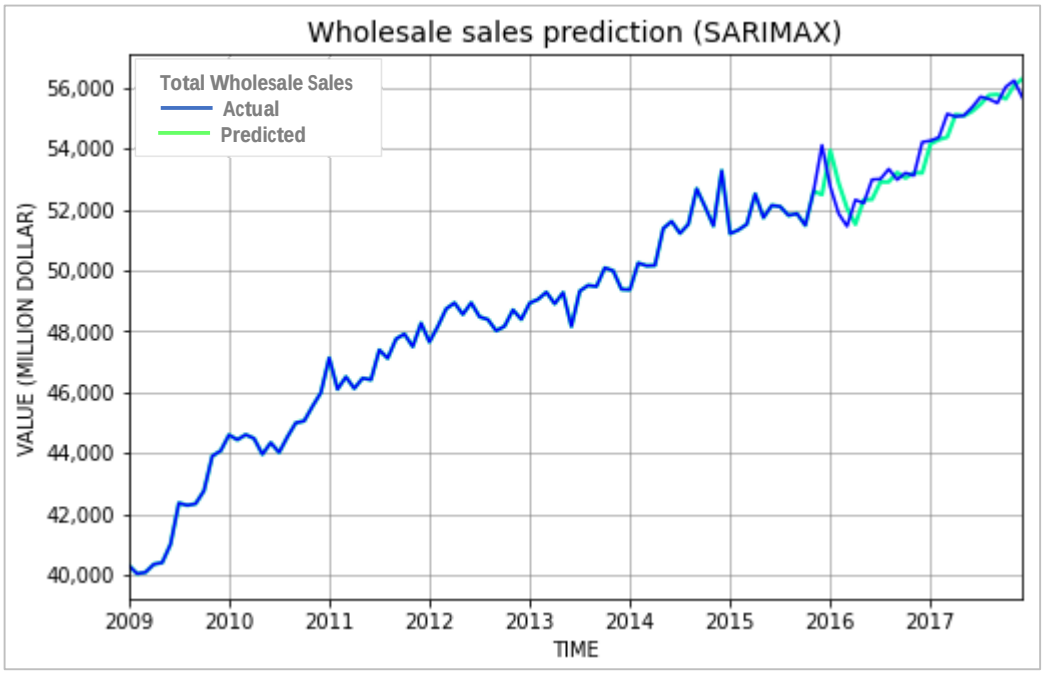


Figure 9 – SARIMAX model prediction for total wholesale sales values

Figure 10 represents the Seasonal Decomposition model prediction for total wholesale sales values applied on train and validation set of data set.

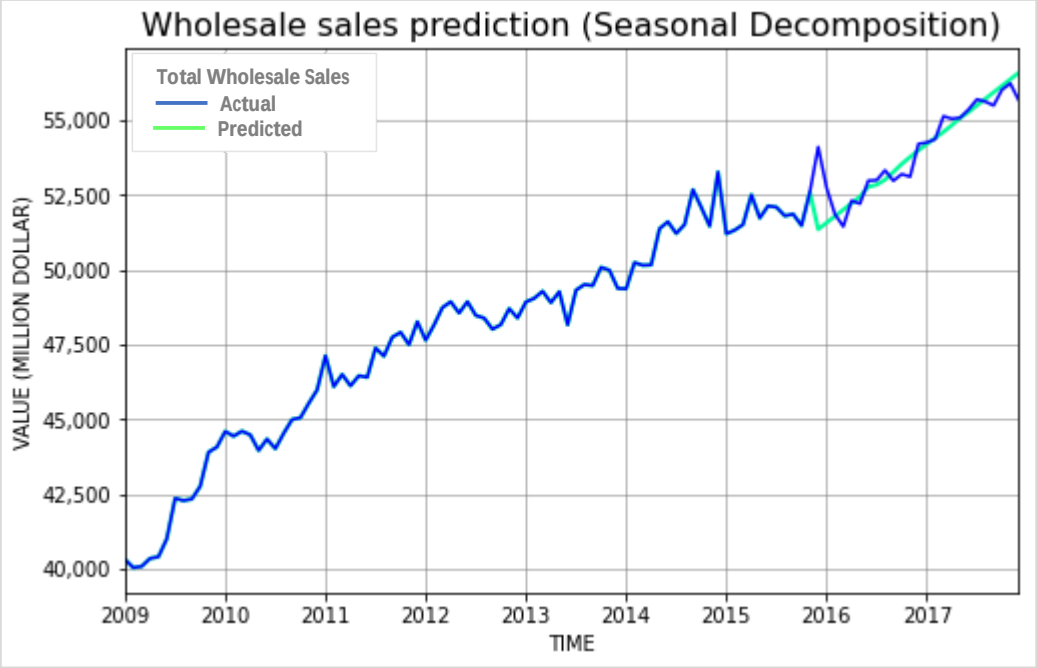


Figure 10 – Seasonal Decomposition model prediction for total wholesale sales values

6.3. Autocorrelation and Partial Autocorrelation Plots

Autocorrelation and partial autocorrelation plots are used in time series analysis and forecasting. These plots are generally used for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. Autocorrelations plots are used in the model identification stage for fittings ARIMA and SARIMAX models. These plots graphically summarize the strength of a relationship with an observation in a time series with observations at prior time steps. According to the plot, the autocorrelations are small valued and not followed the specific pattern for all series. Also, there are positive autocorrelations values within 20 number of lags. If the series has positive autocorrelations within high number of lags and lag-1 autocorrelation is positive valued, then it requires higher order of differencing. For this reason, d (differencing) as 1 gives better result from ARIMA and SARIMAX model implementation instead of d as 0.

Figure 11 and 12 below represents the autocorrelation and partial autocorrelation plot for total wholesale sales data.

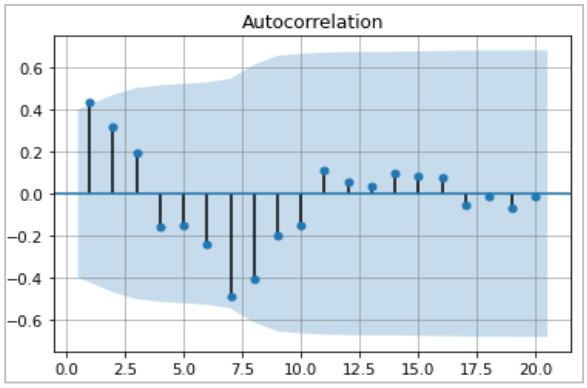


Figure 11 – Autocorrelation plot (total sales)

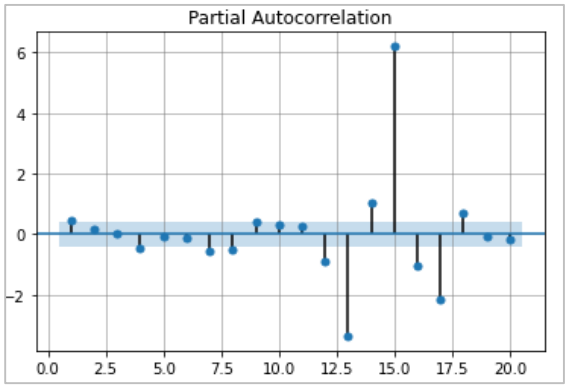


Figure 12 – Partial autocorrelation (total sales)

Figure 13 and 14 below represents the autocorrelation and partial autocorrelation plot for Food merchant wholesalers [4131] which is maximum consumed product of different industry type.

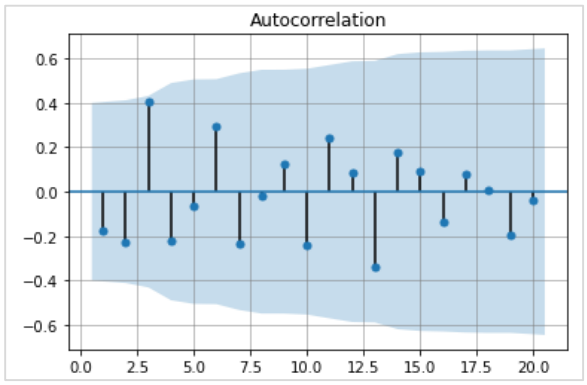


Figure 13 – Autocorrelation plot (4131)

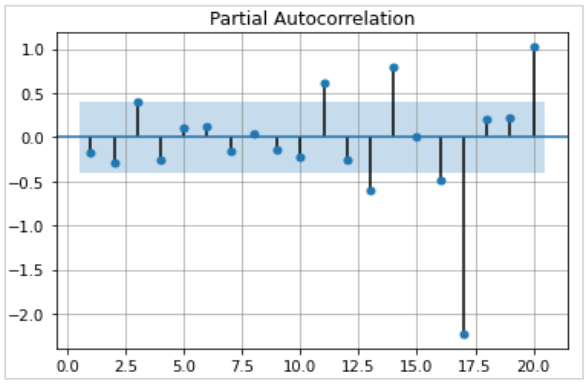


Figure 14 – Partial autocorrelation (4131)

Figure 15 and 16 below represents the autocorrelation and partial autocorrelation plot for Motor vehicle merchant wholesalers [4151] which is 2nd highest consumed product of industry type.

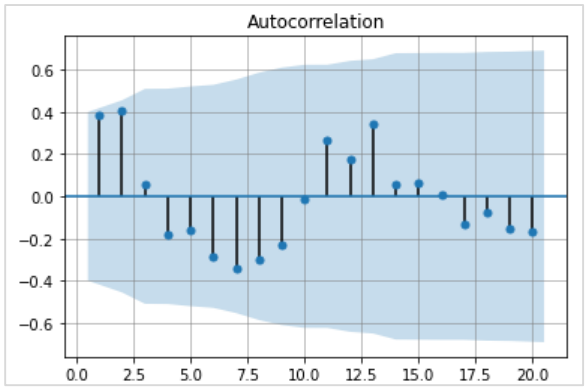


Figure 15 – Autocorrelation plot (4151)

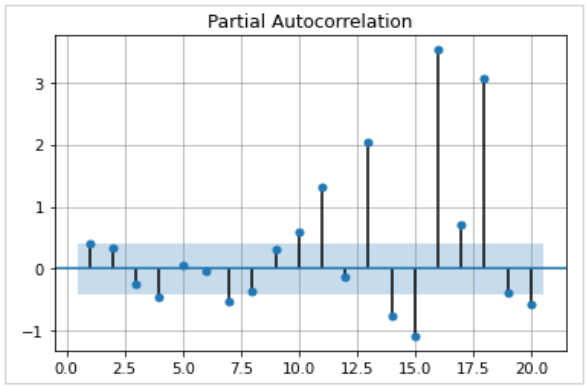


Figure 16 – Partial autocorrelation (4151)

Figure 17 and 18 below represents the autocorrelation and partial autocorrelation plot for Used motor vehicle parts and accessories merchant wholesalers [4153] which is the minimum consumed product of industry type.

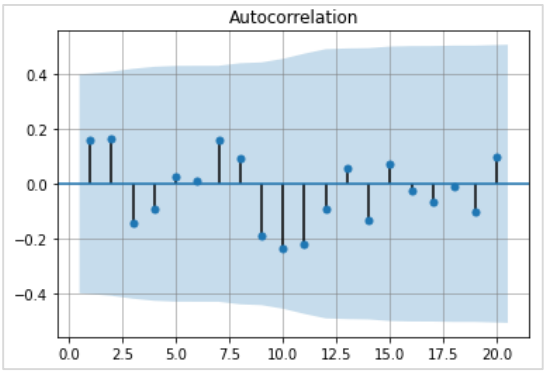


Figure 17 – Autocorrelation plot (4153)

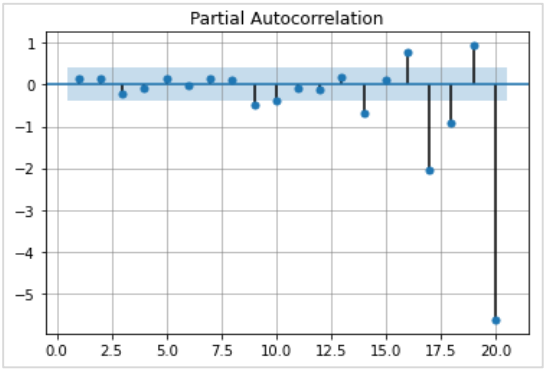


Figure 18 – Partial autocorrelation (4153)

6.4. Experiment Results

a. Experiment 1

In the first experiment of this project, ARIMA Model were applied to predict the data on train and test sets. The hyper parameter settings (p, d, q) as (0, 1, 0) applied which was identified in validation stage of the project. Here, the autoregressive term is zero as the series tends to return to its mean very quickly. In the case of moving average window the value 0 gives better result as there are minimum errors between own lags and lagged forecast errors. The parameters were selected based on minimum MAPE (%) and if we want to get minimum RMSE we need to apply higher order of autoregressive term.

Table 4 represents the summary of ARIMA Model results with parameter (p, d, q) = (0, 1, 0).

Table 4 – ARIMA Model Representation

ARIMA Model Results

Dep. Variable:	D.y	No. Observations:	130
Model:	ARIMA(0, 1, 0)	Log Likelihood	-1017.342
Method:	css	S.D. of innovations	605.934
Date:	Sun, 12 Jul 2020	AIC	2038.684
Time:	03:52:09	BIC	2044.419
Sample:	1	HQIC	2041.015

	coef	std err	z	P>|z|	[0.025	0.975]
const	114.7308	53.144	2.159	0.033	10.571	218.891

	0
count	1.300000e+02
mean	5.361100e-07
std	6.082779e+02
min	-2.181731e+03
25%	-4.059808e+02
50%	2.026923e+01
75%	3.445192e+02
max	1.695269e+03

Figure 19 represents the ARIMA model prediction for total wholesale sales values applied on train and test sets of data set.

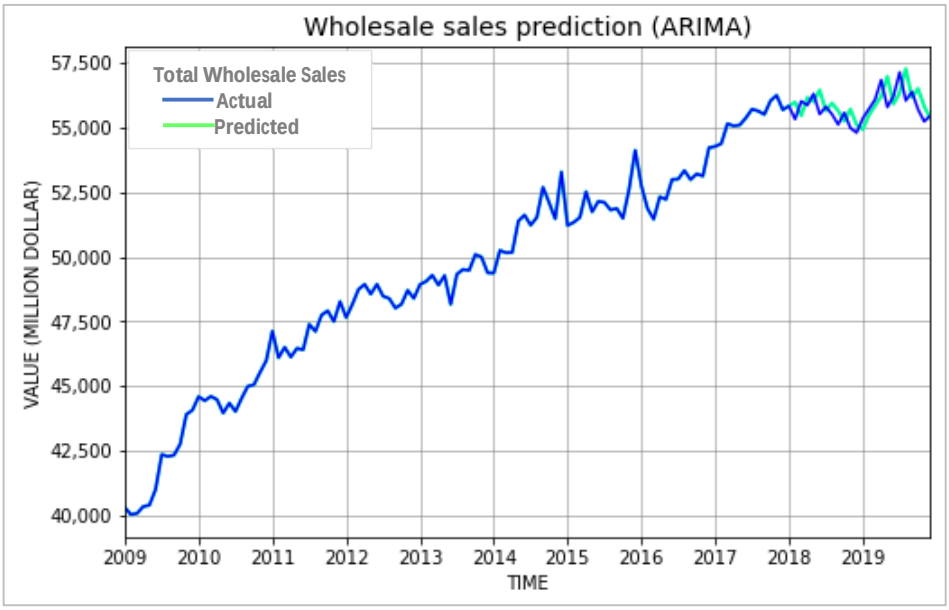


Figure 19 – ARIMA model prediction for total wholesale sales values applied on train/test set

Figure 20 represents the ARIMA model prediction for Maximum consumed product "Food merchant wholesalers [4131]" applied on train and test set of data set.

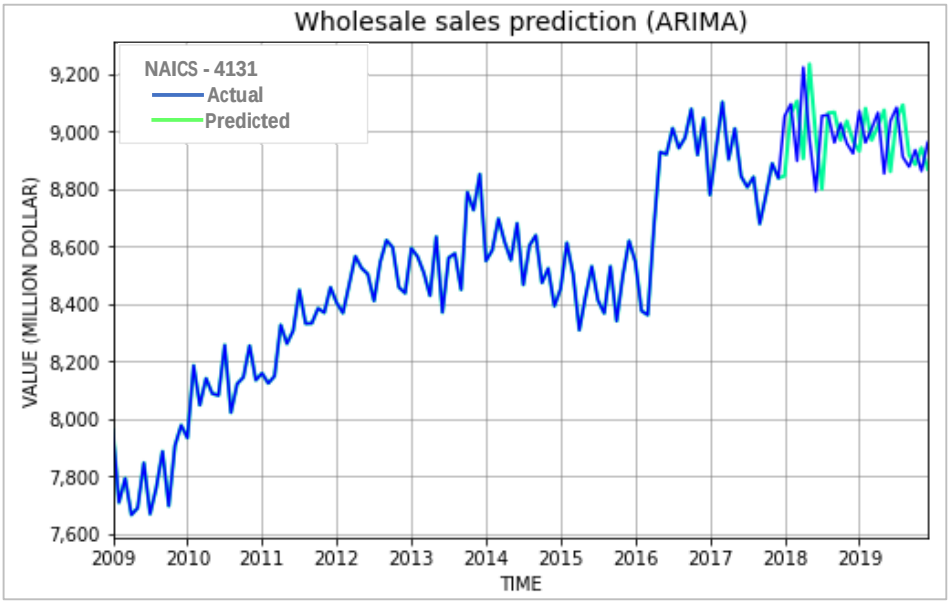


Figure 20 – ARIMA model prediction for Food merchant wholesalers [4131] sales values applied on train/test set

Figure 21 represents the ARIMA model prediction for Maximum consumed product “Motor vehicle merchant wholesalers [4151]” applied on train and test set of data set.

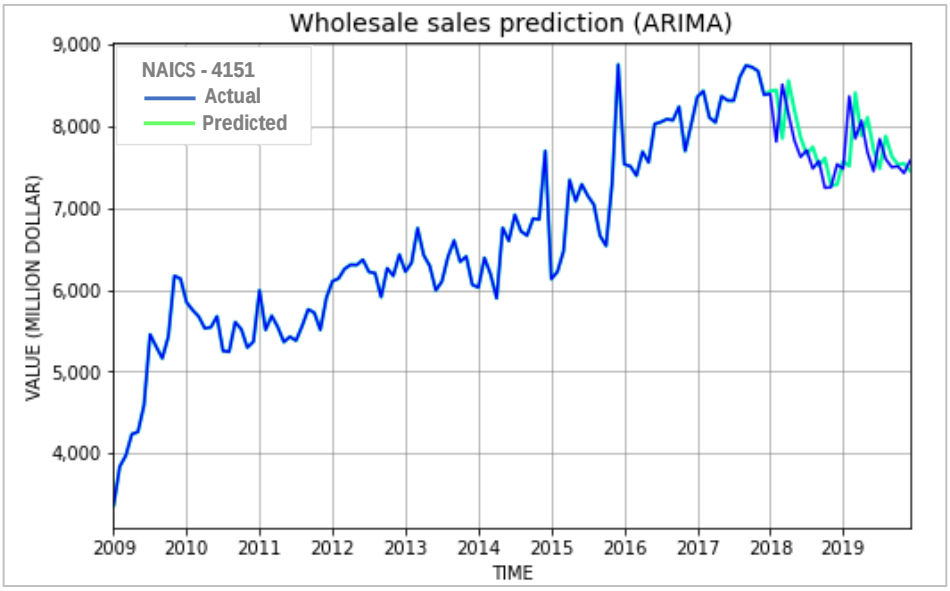


Figure 21 – ARIMA model prediction for Motor vehicle merchant wholesalers [4151] sales values applied on train/test set

Figure 22 represents the ARIMA model prediction for “Used motor vehicle parts and accessories merchant wholesalers [4153]” applied on train and test set of data set.

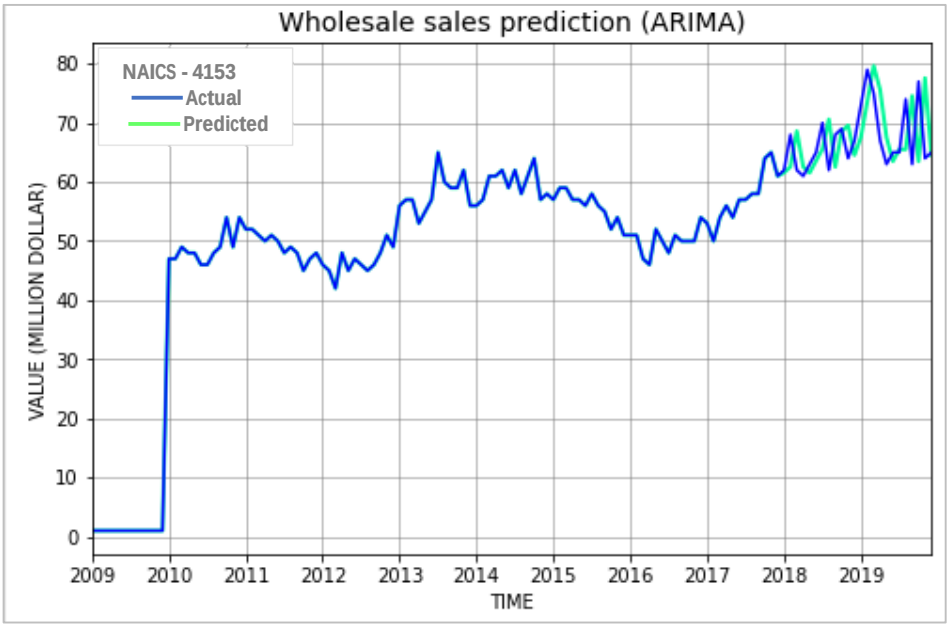


Figure 22 – ARIMA model prediction for Used motor vehicle parts and accessories merchant wholesalers [4153] sales values applied on train/test set

b. Experiment 2

In the second experiment of this project, SARIMAX model were applied to predict the data on train and test sets. The hyper parameter settings (p, d, q) as (6, 1, 1) applied which was identified in validation stage of the project. The hyper parameter settings (p, d, q) as (6, 1, 1) gives minimum RMSE and MAPE (%). According to the result the higher order autoregressive term gives better result as output related to the previous data that are more periods apart. Also, autoregressive model is a random process and specifies that output variables depends linearly on its own previous values. Table 5 represents the summary of SARIMAX Model results with parameters (p, d, q) = (6, 1, 1) applied on train and test set of total wholesale sales values.

Table 5 – SARIMAX Model Representation

```
                           Statespace Model Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  131
Model:               SARIMAX(6, 1, 1)   Log Likelihood               -1011.890
Date:                Sun, 12 Jul 2020   AIC                          2039.779
Time:                        03:52:14   BIC                          2062.720
Sample:                             0   HQIC                         2049.101
                                - 131
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.6178      0.192      3.223      0.001       0.242       0.993
ar.L2          0.0613      0.076      0.802      0.423      -0.089       0.211
ar.L3          0.0381      0.094      0.407      0.684      -0.145       0.221
ar.L4          0.0108      0.083      0.130      0.897      -0.152       0.173
ar.L5          0.0105      0.088      0.119      0.906      -0.162       0.183
ar.L6          0.0949      0.071      1.341      0.180      -0.044       0.234
ma.L1         -0.7526      0.191     -3.950      0.000      -1.126      -0.379
sigma2      3.418e+05   4.26e+04      8.023      0.000    2.58e+05    4.25e+05
===================================================================================
Ljung-Box (Q):                       25.55   Jarque-Bera (JB):                 1.34
Prob(Q):                              0.96   Prob(JB):                         0.51
Heteroskedasticity (H):               1.03   Skew:                            -0.06
Prob(H) (two-sided):                  0.93   Kurtosis:                         3.48
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
                  0
count    131.000000
mean     351.773253
std     3565.379750
min    -1940.909459
25%     -326.790557
50%       52.650975
75%      408.766811
max    40310.000000
```

Figure 23 represents the SARIMAX model prediction for total wholesale values applied on train and test set of data set.

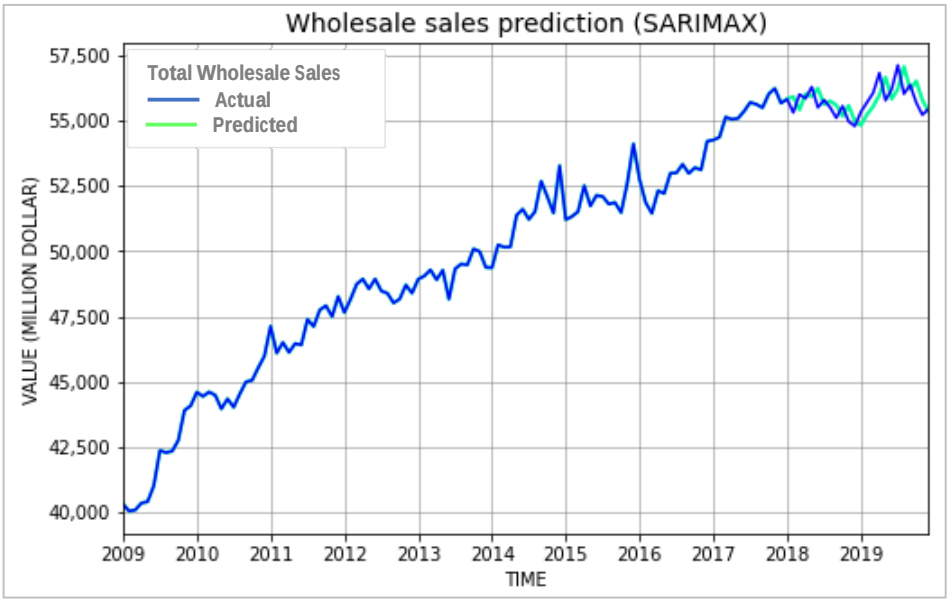


Figure 23 – SARIMAX model prediction for total wholesale sales values applied on train/test set

Figure 24 represents the SARIMAX model prediction for maximum consumed product "Food merchant wholesalers [4131]" applied on train and test set of data set.

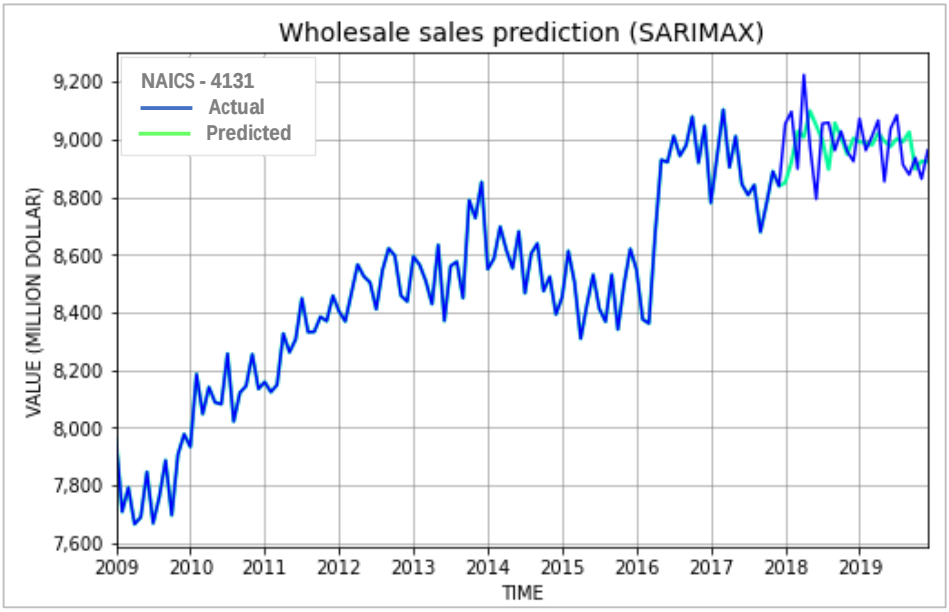


Figure 24 – SARIMAX model prediction for Food merchant wholesalers [4131] sales values applied on train/test set

Figure 25 represents the SARIMAX model prediction for 2nd maximum consumed product “Motor vehicle merchant wholesalers [4151]” applied on train and test set of data set.

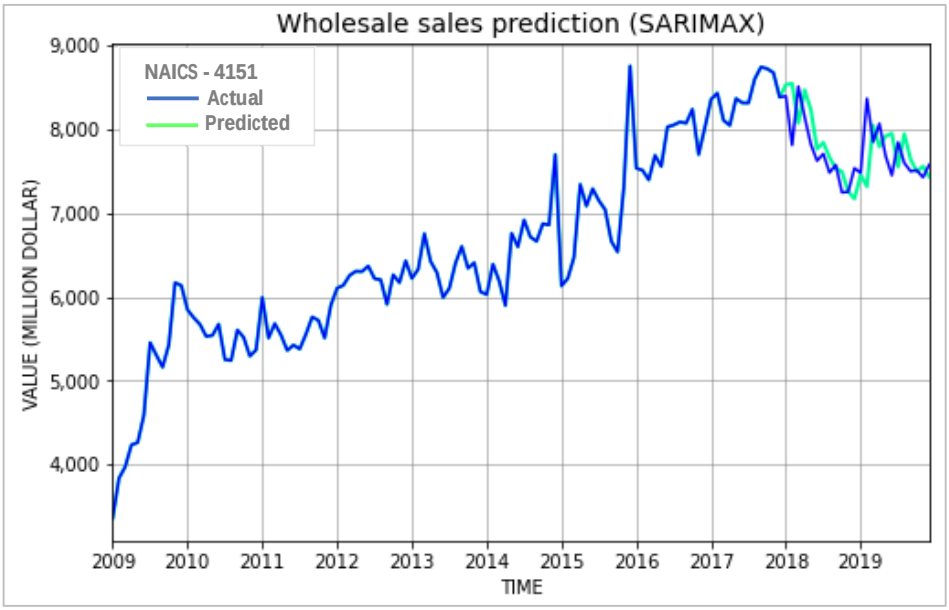


Figure 25 – SARIMAX model prediction for Motor vehicle merchant wholesalers [4151] sales values applied on train/test set

Figure 26 represents the SARIMA model prediction for minimum consumed product “Used motor vehicle parts and accessories merchant wholesalers [4153]” applied on train and test of data set.

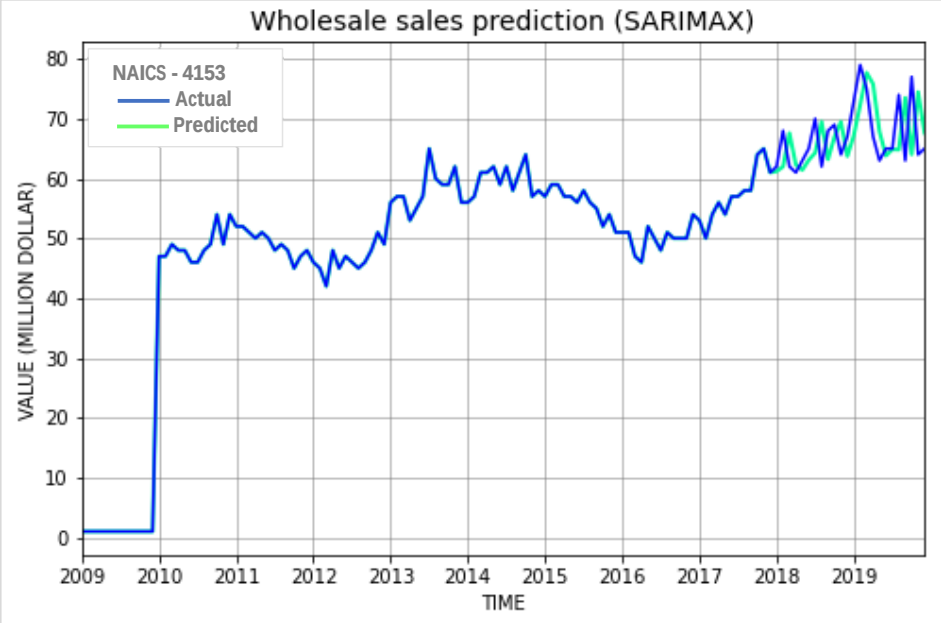


Figure 26 – SARIMAX model prediction for Used motor vehicle parts and accessories merchant wholesalers [4153] sales values applied on train/test set

c. Experiment 3

In the third experiment of this project, Seasonal Decomposition model were applied to get the observed, trained, seasonal and residual data from train and test sets. Finally calculated the RMSE and MAPE (%) between and actual and trend data. As the dataset contain monthly data, the parameter 'freq' set as 12. In this model we can apply either additive or multiplicative settings and only adjust the parameter frequency of the data.

Figure 26 represents the Observed, Trend, Seasonal and Residual part of the Seasonal Decomposition model applied on test sets of data set.

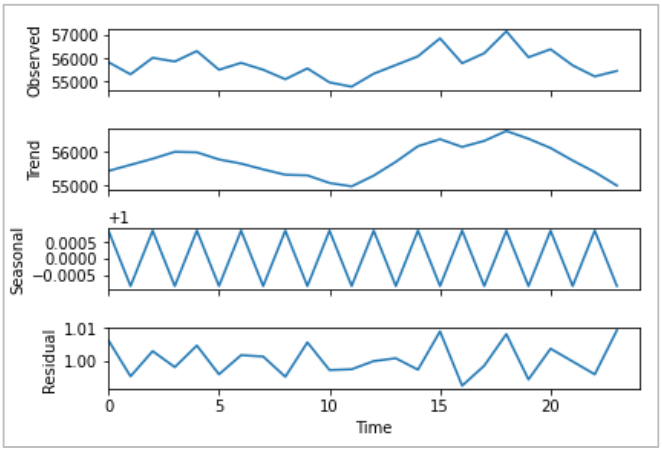


Figure 27 – Observed, Trend. Seasonal and Residual part of Seasonal Decomposition model

Figure 28 represents the Seasonal Decomposition model prediction for total wholesale sales values with parameter 'freq' = 12 applied on test set of data set.

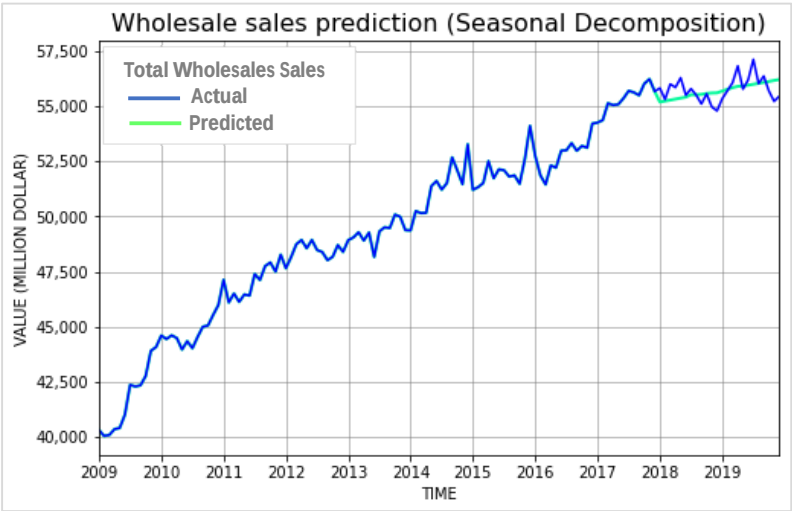


Figure 28 – Seasonal Decomposition model prediction for total wholesale sales values applied on test set of data set

Figure 29 represents the Seasonal Decomposition model prediction for maximum consumed product “Food merchant wholesalers [4131]” applied on test set of data set.

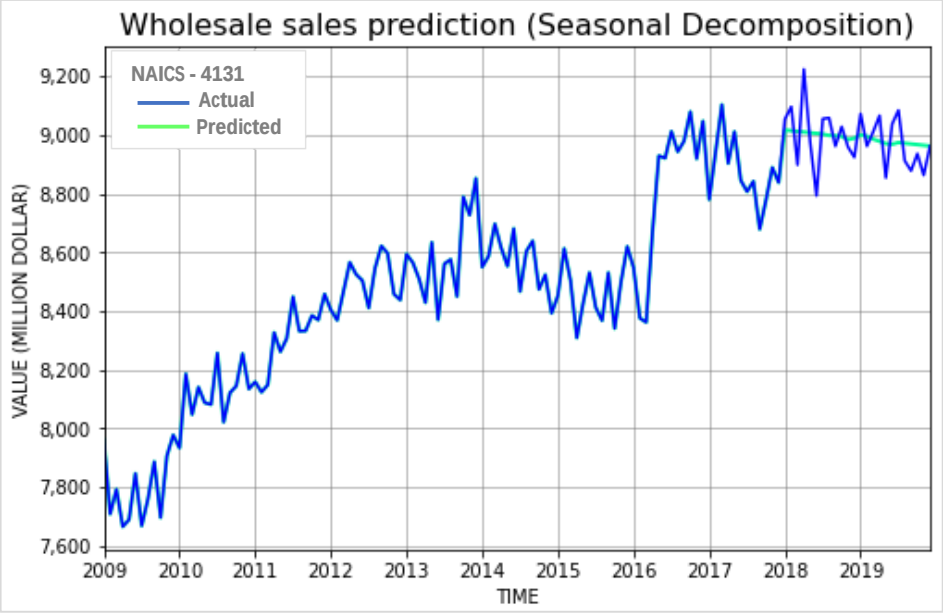


Figure 29 – Seasonal Decomposition model prediction for Food merchant wholesalers [4131] sales values applied on test set of data set

Figure 30 represents the Seasonal Decomposition model prediction for 2nd maximum consumed product “Motor vehicle merchant wholesalers [4151]” applied on test set of data set.

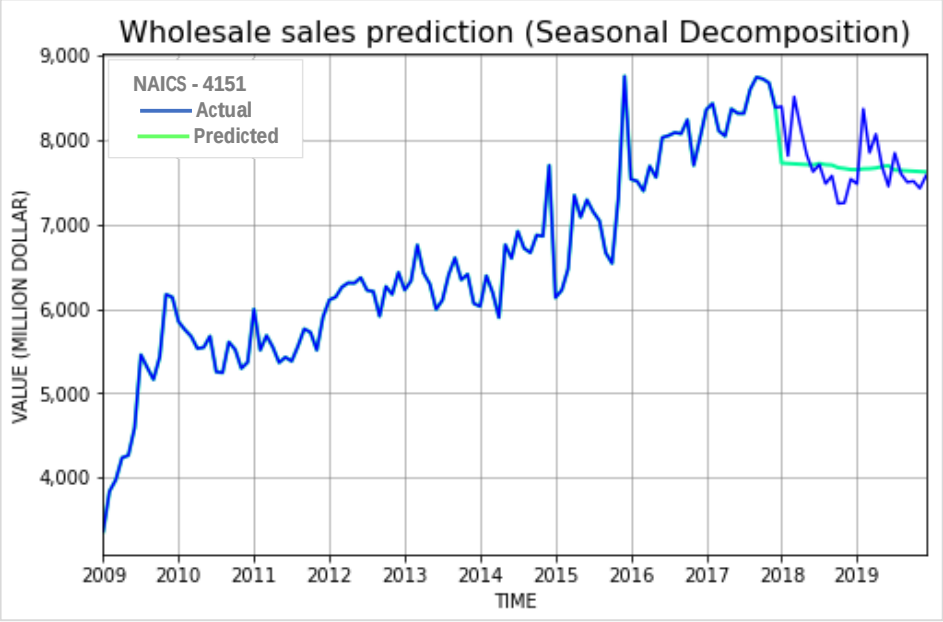


Figure 30 – Seasonal Decomposition model prediction for Motor vehicle merchant wholesalers [4151] sales values applied on test set of data set

Figure 31 represents the Seasonal Decomposition model prediction for minimum consumed product “Used motor vehicle parts and accessories merchant wholesalers [4153]” applied on test set of data set.

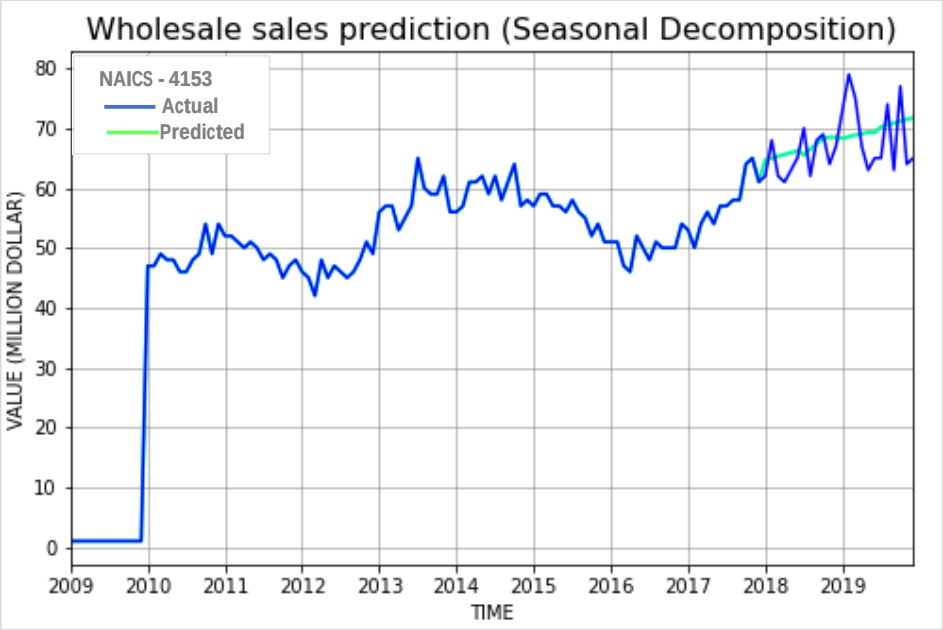


Figure 31 – Seasonal Decomposition model prediction for Used motor vehicle parts and accessories merchant wholesalers [4153] applied on test set of data set

6.5. Prediction Model Performance Result

In this experiment, RMSE and MAPE is used to measure the performance of ARIMA, SARIMAX and Seasonal Decomposition model. For each model, every wholesale sale based on industry type were predicted and measured each of their performance. Based on the result, Seasonal Decomposition gives average minimum RMSE and MAPE, but Seasonal Decomposition not considered the previous values as it only gives trend and residual based on the original trend.

According to the result from ARIMA and SARIMAX, SARIMAX model shows better performing models compared to ARIMA. In the case of average RMSE for all industry product types, SARIMAX shows minimum number of RMSE compared to ARIMA. According to the MAPE result for all industry product types, SARIMAX shows minimum error for each of the product type.

Table 6 represents the measurement of RMSE and MAPE for total wholesale sales based on train and test set with parameter (p, d, q) = (0, 1, 0) for ARIMA, parameter (p, d, q) = (6, 1, 1) for SARIMAX and freq = 12 for Seasonal Decomposition gives minimum RMSE and MAPE errors.

Table 6 – RMSE and MAPE for ARIMA, SARIMAX and Seasonal Decomposition models (Train/Test set)

North American Industry Classification System (NAICS)	ARIMA_RMSE	ARIMA_MAPE	SARIMAX_RMSE	SARIMAX_MAPE	Seasonal Decompose_RMSE	Seasonal Decompose_MAPE
Total Wholesale Sales	591.147605	1.098857	564.137861	0.878346	560.982444	0.797567

Chart 1 represents the RMSE trend from ARIMA, SARIMAX and SD based on each wholesale sales NAICS industry type.

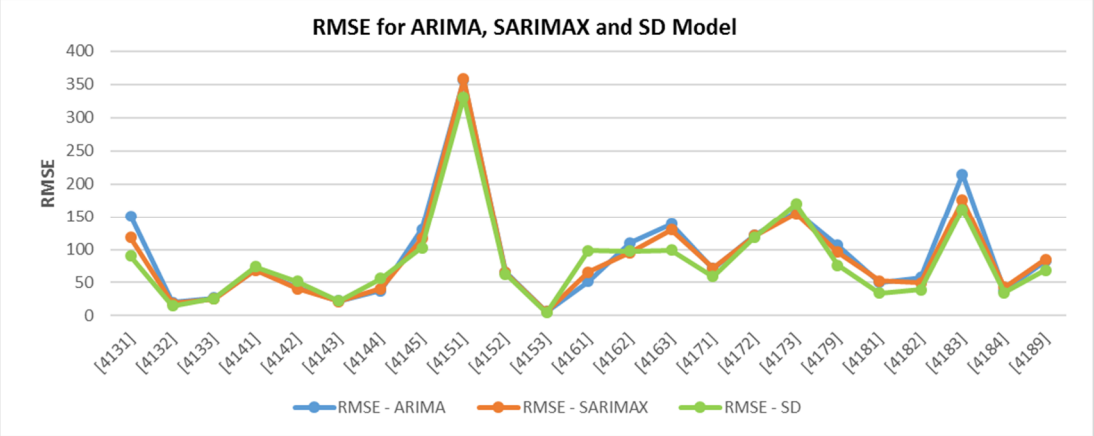


Chart 1 - RMSE trend for ARIMA, SARIMAX and Seasonal Decomposition model

Chart 2 represents the MAPE (%) trend from ARIMA, SARIMAX and SD based on each wholesale sales NAICS industry type.

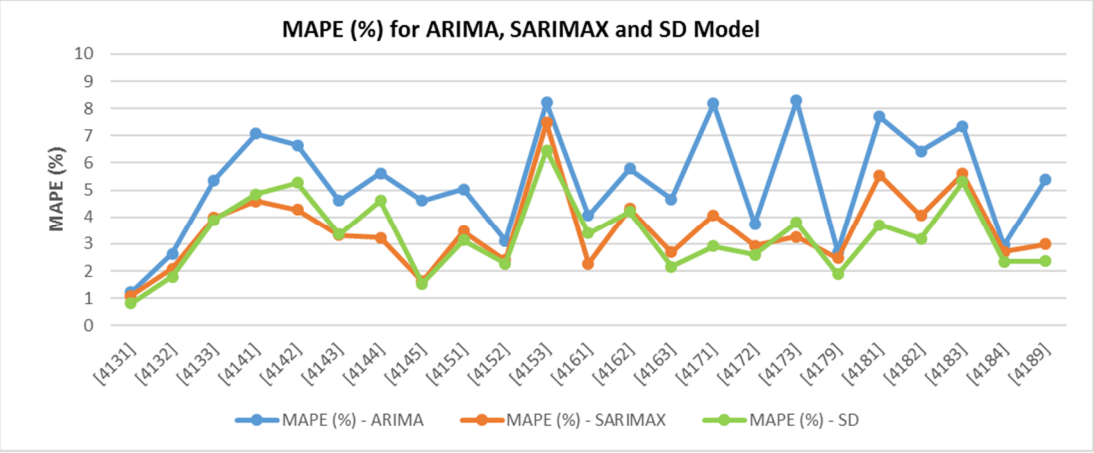


Chart 2 – MAPE (%) trend for ARIMA, SARIMAX and Seasonal Decomposition model

Table 7 represents the RMSE and MAPE for ARIMA, SARIMAX and Seasonal Decomposition Model for each industry type.

Table 7 – RMSE and MAPE for ARIMA, SARIMAX and SD models (Train/Test set)

	RMSE			MAPE		
NAICS	**ARIMA**	**SARIMAX**	**SD**	**ARIMA**	**SARIMAX**	**SD**
[4131]	151.48	118.42	90.02	**1.22**	**1.09**	**0.82**
[4132]	20.38	18.29	15.21	2.63	2.08	1.79
[4133]	26.52	24.99	26.36	5.36	3.99	3.89
[4141]	69.22	68.40	74.10	7.07	4.59	4.85
[4142]	43.17	40.41	52.03	6.65	4.27	5.28
[4143]	22.51	21.87	22.33	4.61	3.32	3.35
[4144]	37.80	40.89	56.06	5.62	3.22	4.61
[4145]	130.64	116.75	102.93	4.60	1.63	1.54
[4151]	**356.83**	**358.26**	**330.99**	5.04	3.46	3.14
[4152]	65.46	65.07	62.97	3.11	2.40	2.27
[4153]	**6.43**	**6.12**	**4.95**	8.22	**7.48**	**6.46**
[4161]	51.93	65.43	97.97	4.06	2.27	3.40
[4162]	110.00	95.25	97.65	5.78	4.31	4.21
[4163]	138.93	129.83	98.71	4.67	2.69	2.16
[4171]	71.08	71.38	58.80	**8.19**	4.08	2.92
[4172]	121.80	121.22	118.76	3.74	2.94	2.60
[4173]	156.89	153.80	169.52	8.29	3.26	3.80
[4179]	106.12	96.21	76.21	2.72	2.47	1.88
[4181]	50.58	52.26	33.79	7.71	5.54	3.71
[4182]	57.88	49.26	39.59	6.43	4.06	3.19
[4183]	214.31	175.53	161.57	7.35	5.61	5.31
[4184]	37.53	42.79	35.08	2.96	2.72	2.35
[4189]	82.91	84.79	68.54	5.40	2.98	2.37
Average	92.63	87.71	82.35	5.28	3.50	3.30

Density Plot: Probability density is the relationship between observations and their probability. A density plot is a representation of the distribution of one or a few numeric variables and it uses a kernel density estimate to represent the probability density function of the variable and can be interpreted as providing a relative likelihood that the value of the random variable.

Figure 32 and 33 represents the density plot of ARIMA and SARIMAX for total wholesales data.

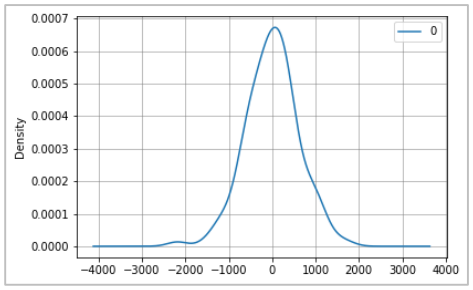


Figure 32 – ARIMA for total wholesales data

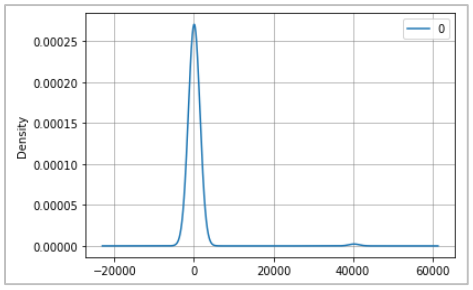


Figure 33 – SARIMAX for total wholesales data

Figure 34 represents the residual plot of ARIMA for total wholesales data.

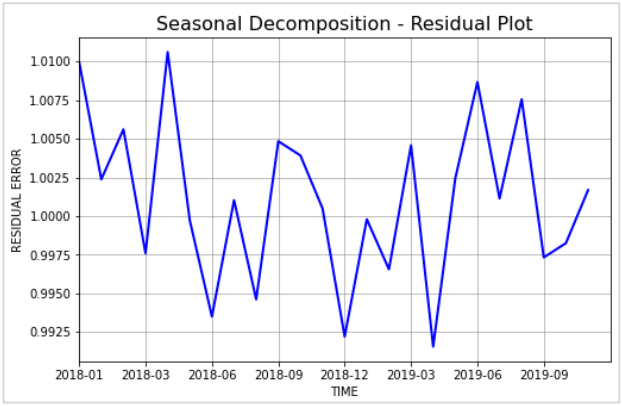


Figure 34 – Residual plot of ARIMA for total wholesales data.

Figure 35 and 36 represents the density plot of ARIMA and SARIMAX for Food merchant wholesalers [4131] which is maximum consumed product.

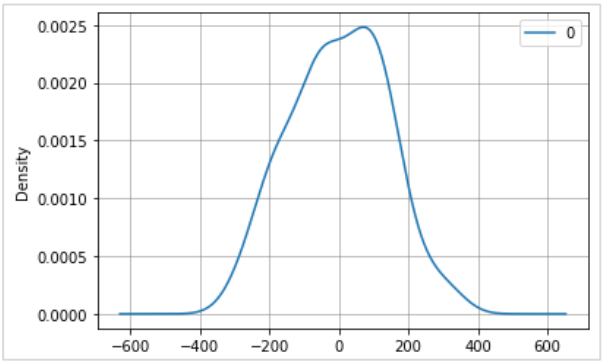


Figure 35 – Density plot of ARIMA for 4131

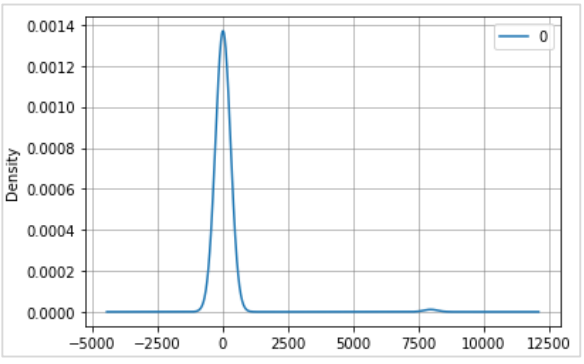


Figure 36 – Density plot of SARIMAX for 4131

Figure 37 represents the residual plot of Seasonal Decomposition for Food merchant wholesalers [4131] which is maximum consumed product.

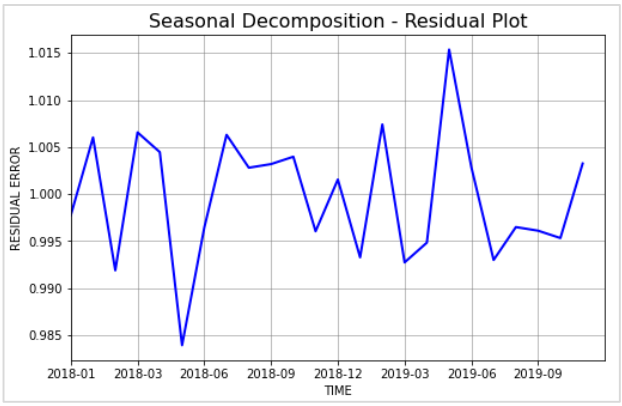
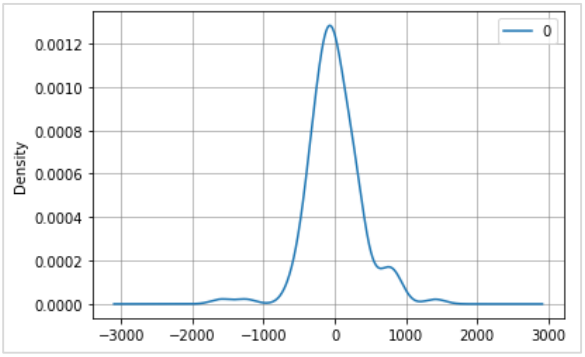


Figure 37 – Residual plot of Seasonal Decomposition for Food merchant wholesalers [4131]

Figure 38 and 39 represents the density plot of ARIMA and SARIMAX for Food merchant wholesalers [4151] which is 2nd maximum consumed product.



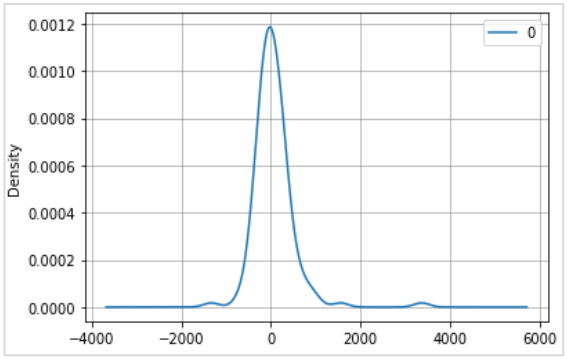


Figure 38 – Density plot of ARIMA for [4151]

Figure 39 - Density plot of SARIMAX for [4151]

Figure 40 represents the residual plot of Seasonal Decomposition for Food merchant wholesalers [4151] which is 2nd maximum consumed product.

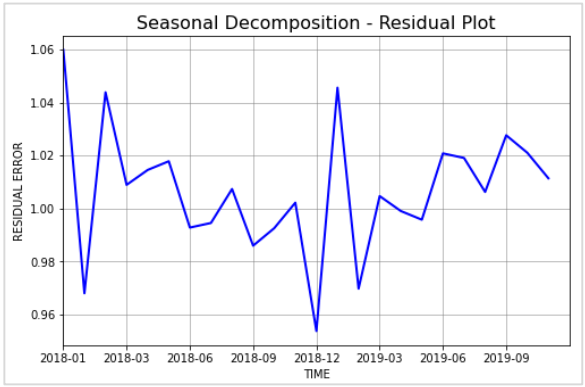
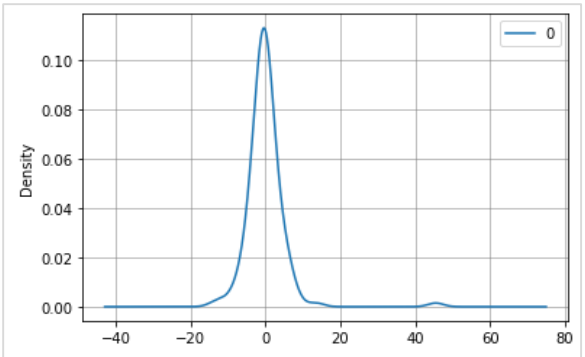


Figure 40 – Residual plot of Seasonal Decomposition for Food merchant wholesalers [4151]

Figure 41 and 42 and represents the density plot of ARIMA and SARIMAX for Used motor vehicle parts and accessories merchant wholesalers [4153] which is minimum consumed product.



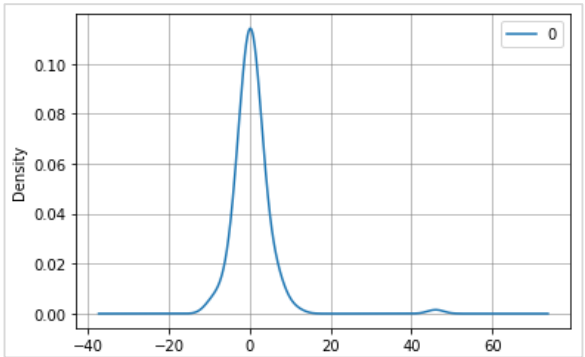


Figure 41 – Density plot of ARIMA for [4153]

Figure 42 – Density plot of SARIMAX for [4153]

Figure 43 represents the residual plot of Seasonal Decomposition for Used motor vehicle parts and accessories merchant wholesalers [4153] which is minimum consumed product.

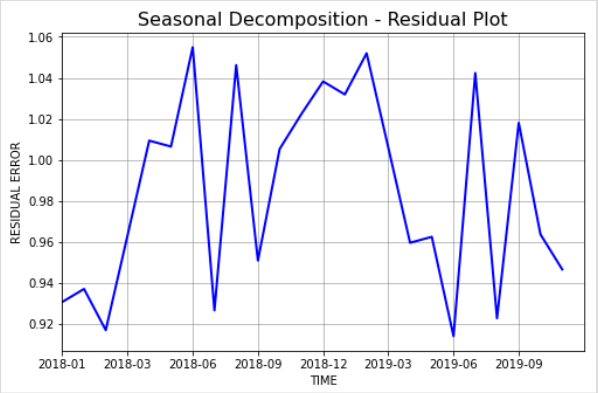


Figure 43 – Residual plot of Seasonal Decomposition for Used motor vehicle parts and accessories merchant wholesalers [4153]

6.6. Discussion

The results of this project have interesting implications for implementing different model in wholesale sales prediction and calculate errors between actual and forecast results. In this project three forecasting model were implemented and found that SARIMAX is the best that gives minimum RMSE and MAPE (%). If we compare MAPE (%) between ARIMA and SARIMAX, SARIMAX model gives the minimum errors for all wholesale sales product type. Also, efforts were made to calculate errors for different combination of hyper-parameters for ARIMA and SARIMAX model and selected best parameter settings that gives minimum errors between actual and predicated data.

Seasonal Decomposition model helps to decompose the time series data into different components which helps to determine the suitable prediction model as decomposition is more for analysis than prediction. After decomposed the result it is required to fit a model to make a prediction. In this project ARIMA and SARIMAX model used which gives minimum RMSE and MAPE (%). The forecasting model ARIMA and SARIMAX considers the previous values to calculate the prediction values but Seasonal Decomposition model has no feature to consider the previous values. In the case of decomposed the data, Seasonal Decomposition model gives the good idea to represent the residual errors, trend and calculate the errors between observed and trend components of the model.

In this project, three forecasting models were implemented for NAICS wholesale sales data and SARIMAX is the best forecasting model as it considers the previous values to predict the future values and it has feature to apply external values as parameter.

7. CONCLUSION AND FUTURE WORKS

In this research project, three forecasting models were implemented to predict wholesale sales product demand based on wholesale sales time series data. According to the prediction result and calculated errors, SARIMAX is the best model to predict the wholesale sales data. Seasonal Decomposition model gives the minimum number of errors, but this model has no feature to consider previous values.

The dataset that used in this project reflect wholesale sales based on different industry type that helped to get the summaries about different prediction values and calculated errors. But the dataset did not contain enough historical data, however if forecasting models that used in this project will be implemented for large amount of daily transaction data then this would be a potential forecasting models for large business transactions. In order to generate the reliable forecast for decision making in real business environment, sometimes required to develop a system that can deal with the automatic parameterisation of forecasting models and select appropriate forecasting techniques from set of forecasting models in the system.

The demand forecasting with time series data is a fast-growing area of research and as such provides many scopes for future works. Firstly, in the case of Seasonal Decomposition model, if we can integrate the parameter to consider previous trend values to predict the future values, then this model will be a good baseline for forecasting wholesale sales values compared to other forecasting model as it gave minimum error between observed and trend data. Secondly, the forecasting models were implemented independently without considered merging or ensemble methods. Potential future research therefore includes an ensemble method to combine different forecasting techniques and minimize the errors between actual and predicted data. Finally, the future work will anticipate geographical location, population and weather as external attributes to build the inimitable model that can help to forecast the specific type of time series data.

APPENDIX – A: GITHUB LINK

Github Link for MRP Data Set and Python Code

https://github.com/wimurad/DS_MRP_2020

APPENDIX – B: LIST OF FIELDS IN THE DATASET

[1] REF_DATE

[2] GEO

[3] DGUID

[4] Sales, price and volume

[5] North American Industry Classification System (NAICS)

[6] UOM

[7] UOM_ID

[8] SCALAR_FACTOR

[9] SCALAR_ID

[10] VECTOR

[11] COORDINATE

[12] VALUE

[13] STATUS

[14] SYMBOL

[15] TERMINATED

[16] DECIMALS

[17] YEAR

REFERENCES

Abbasimehr, H., Shabani, M., & Yousefi, M. (2020). An optimized model using LSTM network for demand forecasting. Computers & Industrial Engineering, 106435.

Abolghasemi, M., Beh, E., Tarr, G., & Gerlach, R. (2020). Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion. *Computers & Industrial Engineering*, 106380.

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019, December). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *International Conference on Neural Information Processing* (pp. 462-474). Springer, Cham.

Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2019). Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting*, *35*(1), 170-180.

Bruzda, J. (2019). Quantile smoothing in supply chain and logistics forecasting. *International Journal of Production Economics*, *208*, 122-139.

Chen, C., Liu, Z., Zhou, J., Li, X., Qi, Y., Jiao, Y., & Zhong, X. (2019, April). How Much Can A Retailer Sell? Sales Forecasting on Tmall. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 204-216). Springer, Cham.

Hasni, M., Aguir, M. S., Babai, M. Z., & Jemai, Z. (2019). On the performance of adjusted bootstrapping methods for intermittent demand forecasting. *International Journal of Production Economics*, *216*, 145-153.

Hirt, R., Kühl, N., Peker, Y., & Satzger, G. (2020). How to Learn from Others: Transfer Machine Learning with Additive Regression Models to Improve Sales Forecasting. *arXiv preprint arXiv:2005.10698*.

Jiao, X., Li, G., & Chen, J. L. (2020). Forecasting international tourism demand: a local spatiotemporal model. Annals of Tourism Research, 83, 102937.

Karb, T., Kühl, N., Hirt, R., & Glivici-Cotruta, V. (2020). A network-based transfer learning approach to improve sales forecasting of new products. *arXiv preprint arXiv:2005.06978*.

Kazemzadeh, M. R., Amjadian, A., & Amraee, T. (2020). A hybrid data mining driven algorithm for long term electric peak load and energy demand forecasting. *Energy*, 117948.

Kechyn, G., Yu, L., Zang, Y., & Kechyn, S. (2018). Sales forecasting using WaveNet within the framework of the Kaggle competition. *arXiv preprint arXiv:1803.04037*.

Lee, D., Jung, S., Cheon, Y., Kim, D., & You, S. (2019). Demand Forecasting from Spatiotemporal Data with Graph Networks and Temporal-Guided Embedding. *arXiv preprint arXiv:1905.10709*.

Lin, Z., Madotto, A., Winata, G. I., Liu, Z., Xu, Y., Gao, C., & Fung, P. (2019). Learning to Learn Sales Prediction with Social Media Sentiment. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing* (pp. 47-53).

Pereira, M. M., Machado, R. L., Pires, S. R. I., Dantas, M. J. P., Zaluski, P. R., & Frazzon, E. M. (2018). Forecasting scrap tires returns in closed-loop supply chains in Brazil. *Journal of Cleaner Production*, *188*, 741-750.

Rivera, R., & Burnaev, E. (2017, November). Forecasting of commercial sales with large scale Gaussian Processes. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 625-634). IEEE.

Rivera-Castro, R., Nazarov, I., Xiang, Y., Pletneev, A., Maksimov, I., & Burnaev, E. (2019, July). Demand forecasting techniques for build-to-order lean manufacturing supply chains. In *International Symposium on Neural Networks* (pp. 213-222). Springer, Cham.

Roque, L., Fernandes, C. A., & Silva, T. (2019). Optimal Combination Forecasts on Retail Multi-Dimensional Sales Data. *arXiv preprint arXiv:1903.09478*.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, *252*(1), 1-26.

Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2019). Demand forecasting in restaurants using machine learning and statistical analysis. *Procedia CIRP*, *79*, 679-683.

Van Belle, J., Guns, T., & Verbeke, W. (2020). Using shared sell-through data to forecast wholesaler demand in multi-echelon supply chains. European Journal of Operational Research.

Van Calster, T., Bossche, F. V. D., Baesens, B., & Lemahieu, W. (2020). Profit-oriented sales forecasting: a comparison of forecasting techniques from a business perspective. *arXiv preprint arXiv:2002.00949*.

Wang, C. H., & Chen, J. Y. (2019). Demand forecasting and financial estimation considering the interactive dynamics of semiconductor supply-chain companies. *Computers & Industrial Engineering*, *138*, 106104.

Xu, Q., & Sharma, V. (2017, July). Ensemble Sales Forecasting Study in Semiconductor Industry. In *Industrial Conference on Data Mining* (pp. 31-44). Springer, Cham.

Zhao, K., & Wang, C. (2017). Sales Forecast in E-commerce using Convolutional Neural Network. *arXiv preprint arXiv:1708.07946*.