Master of Science in Computer Science
October 2019

Machine Learning Approach for Forecasting the Sales of Truck Components

Venishetty Sai Vineeth

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfilment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Author(s):
Venishetty Sai Vineeth
E-mail: vesi17@student.bth.se

University advisor:
Dr. Huseyin Kusetogullari
Department of Computer Science

Faculty of Computing
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Abstract

Context: The context of this research is to forecast the sales of truck components using machine learning algorithms that play an important role in optimizing truck sales process in addressing issues such as delivery time, stock maintenance, marketing, and discounts, etc and also plays a major role in decision-making operations in the areas corresponding to sales, production, purchasing, finance and accounting.

Objectives: This study first investigates to find the suitable machine learning algorithms that can be used to forecast the sales of truck components and then the experiment is performed with the chosen algorithms to forecast the sales and to evaluate the performances of the chosen machine learning algorithms.

Methods: Firstly, a Literature review is used to find suitable machine learning algorithms and then based on the results obtained, an experiment is performed to evaluate the performances of machine learning algorithms.

Results: Results from the literature review shown that regression algorithms namely Supports Vector Machine Regression, Ridge Regression, Gradient Boosting Regression, and Random Forest Regression are suitable algorithms and results from the experiment showed that Ridge Regression has performed well than the other machine learning algorithms for the chosen dataset.

Conclusion: After the experimentation and the analysis, the Ridge regression algorithm has been performed well when compared with the performances of the other algorithms and therefore, Ridge Regression is chosen as the optimal algorithm for performing the sales forecasting of truck components for the chosen data.

Keywords: Machine Learning, Time Series Forecasting, Sales Forecasting.

Acknowledgments

I would first like to express my deep sense of gratitude and thanks to Dr. Huseyin Kusetogullari for his exceptional guidance, supervision and encouragement. I would also like to thank supervisors at Volvo Group, Alain Boone and Nina Xiangni Chang, who guided and helped me in understanding the real time problem statement and in finding an optimal solution.

Finally, I would like to thank my parents and friends for their tremendous support and encouragement.

Contents

Abstract i

Acknowledgments ii

1 Introduction 1
- 1.1 Aim and Objectives . . . 3
- 1.2 Research Questions . . . 3

2 Background 4
- 2.1 Time Series . . . 4
  - 2.1.1 Univariate Time Series . . . 5
  - 2.1.2 Multivariate Time Series . . . 5
- 2.2 Machine Learning . . . 6
- 2.3 Machine Learning Algorithms . . . 7
  - 2.3.1 Random Forest Regression . . . 7
  - 2.3.2 Support Vector Regression . . . 7
  - 2.3.3 Ridge Regression . . . 8
  - 2.3.4 Gradient Boosting Regression . . . 8
  - 2.3.5 Selection of Algorithms . . . 9

3 Related Work 10

4 Method 13
- 4.1 Literature Review . . . 13
- 4.2 Experiment . . . 13
- 4.3 Software Environment . . . 15
  - 4.3.1 Python . . . 15
  - 4.3.2 Jupyter Notebook . . . 16
- 4.4 Dataset description . . . 16
- 4.5 Data Preprocessing . . . 16
  - 4.5.1 Categorical Encoding . . . 16
  - 4.5.2 Sliding window method . . . 18
  - 4.5.3 Augmented Dickey-Fuller test . . . 18
  - 4.5.4 Normalization . . . 19
- 4.6 Feature Selection . . . 20
  - 4.6.1 Correlation method . . . 20
- 4.7 Experimental Setup . . . 21
  - 4.7.1 Performance Metrics . . . 21

4.8 Time Series Cross-Validation 22

5 Results **24**
5.1 Stationarity Check 24
5.2 Gradient Boosting Regressor 24
5.3 Random Forest Regressor 25
5.4 Support Vector Regressor (SVR) 26
5.5 Ridge Regressor 28
5.6 Performance evaluation results 29

6 Analysis and Discussion **30**
6.1 Comparative study of Performance Metrics 30
6.1.1 Mean Absolute Error 30
6.1.2 Root Mean Squared Error 30
6.1.3 Average Mean Absolute Error 31
6.1.4 Average Root Mean Square Error 32
6.2 Key Analysis 32
6.3 Discussion 33
6.4 Contributions 33
6.5 Threats to Validity 33
6.5.1 Internal Validity 33
6.5.2 External Validity 34
6.5.3 Conclusion Validity 34
6.6 Limitations 34

7 Conclusions and Future Work **35**
7.1 Future Work 35

List of Figures

1.1 Sales Forecasting in Business Process 2

2.1 Univariate Time Series 5
2.2 Multi-Variate Time Series 6
2.3 Support Vector Regression 8

4.1 Mechanism followed in this thesis 14
4.2 One hot encoding 17
4.3 Sample Time Series Data 18
4.4 After applying sliding window 18
4.6 Correlation with the target Variant 21
4.7 Time Series Cross Validation 23

5.1 Augmented Dickey-Fuller Test 24
5.2 Gradient Boosting Error Box Plots 25
5.3 Gradient Boosting Regression Prediction 25
5.4 Random Forest Regression Error Box Plots 26
5.5 Random Forest Regressor Prediction 26
5.6 Support Vector Regression Error Box Plots 27
5.7 Support Vector Regressor Prediction 27
5.8 Ridge Regression Error Box Plots 28
5.9 Ridge Regressor Prediction 28

6.1 Comparison of MAE obtained by regression models on 5-fold time series validation tests 30
6.2 Comparison of RMSE obtained by regression models on 5-fold time series validation tests 31
6.3 Comparison of Average MAE obtained by regression models on 5-fold time series validation tests 31
6.4 Comparison of Average RMSE obtained by regression models on 5-fold time series validation tests 32

List of Tables

3.1 Summarization of the literature review 12

5.1 Comparison of performance evaluation results 29

Chapter 1

Introduction

Good forecasts play a vital role in many fields of scientific, industrial, commercial and economic activity [1]. In today's business world's consumer-centric environment, companies seeking good sales performance often need to maintain a balance between meeting customer demand and controlling cost of inventory. Carrying a bigger inventory enables client demand to be satisfied at all times, but may result in over-stocking, leading to issues such as tied-up capital, written down inventory, and lower profit margins. In comparison, lower inventory concentrations may decrease inventory expenses, but may result in cost of chance resulting from missed selling possibilities, lower customer satisfaction, and other issues. Sales forecasting is the process of determining future sales and the forecasts can be used to maintain the necessary amount of inventory to prevent the under or over-stocking issues. Sales forecasting can affect corporate financial planning, marketing, customer management, and other company fields. Consequently, improving the precision of sales forecasts has become a significant element of a company operation [2].

Sales forecasting is a more traditional but still very compelling application of time series forecasting [3]. Time series forecasting is being used as the foundation for the functioning of any process over the time based on the past data. Forecasts are determined by using the data from the past and by considering the known factors in the future [4]. Much effort is dedicated over the past decades for the development and improvement of various forecasting models. The characteristics of time series are essentially noisy, non-stationary, non-linear, unstructured along with the many influencing factors of political, economic and psychological identity made many applications related to exchange rate as difficult applications of financial forecasting techniques [5].

In a survey conducted by McKinsey Global Institute on adoption and use of Artificial intelligence across various sectors show that the financial services sector is leading [6]. Sales forecasting plays a major role in financial planning and conducting business for any organization to assess the statistics of the past and current sales and to forecast future performance. Altogether, precise sales forecasting helps the company to run more productively and efficiently, to save money on the approaches to make forecasts or predictions described as Statistical Modelling, Machine learning [7].

There are about 12,000 individual parts of a truck with different specifications, where some parts are assembled with robots and most of the parts are assembled manually at various workstations. A customer can customize a truck by choosing around eleven thousand different parts that make it about 80-90 percent of a truck. Customers select different parts depending on the various aspects like region, variation in size, etc. It is therefore difficult to determine which components the customers end up choosing. Thus, assessing sales of the components with the use of forecasting plays an important role in optimizing the truck sales process of organizations in overcoming issues like delivery time, stock maintenance, marketing, and discounts, etc. As the parts are large in number, they have been categorized into families like Core Components, Rim and Types, Cab interior-driving, etc.

This thesis uses machine learning [8] approach to forecast the sales of the parts related to the family of the core components of a truck, which consists of the Transport segment, Chassis Height, front axle load etc. By using the machine learning approach, obtained sales forecast results could help the organization to assess the sales of the components of the truck, to maintain stock of the components which have more sales, saves money and time from manufacturing the items which have least sales or no sales in a particular region. The dataset used in this thesis consists of the data related to the sales of core components of Volvo Trucks obtained from the Digital Transformation department of Volvo group.

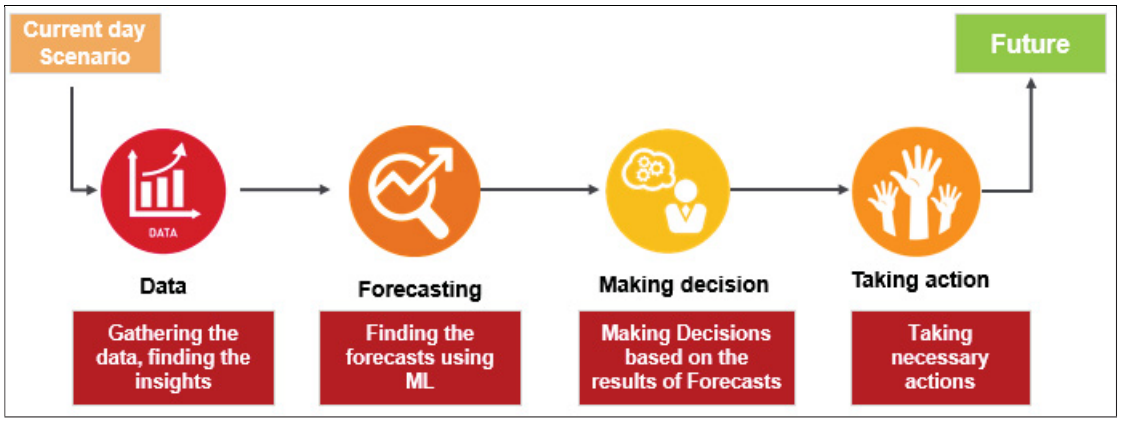


Figure 1.1: Sales Forecasting in Business Process

1.1 Aim and Objectives

The aim of this thesis is to investigate the various sales forecasting methods executed in financial area and evaluate the performance of the chosen machine learning algorithms to find out the best suitable and efficient model for the chosen data set.

Objectives

- To understand the efficient machine learning techniques for forecasting the sales.
- To evaluate the performance of the selected machine learning algorithms by comparing the degree of prediction success rate.

1.2 Research Questions

Two research questions have been formulated for this study to successfully accomplish the aim are as follows:

RQ1: What are the suitable machine learning techniques for forecasting the sales in the area of finance?

Motivation: The motivation of this research question is to study the appropriate machine learning algorithms used for forecasting the sales in financial area.

RQ2: What are the performances of machine learning models for forecasting the sales of the core components of trucks based on the historical data?

Motivation: The motivation of this research question is to find out the efficient machine learning algorithm among the chosen models after comparison of results based on the performance evaluation metrics.

The appropriate machine learning algorithms for sales forecasting are obtained from the literature review is selected to answer RQ1. The chosen machine learning models are analyzed, and the efficient machine learning algorithm is selected based on the results obtained after the performance evaluation answers RQ2. Performance metrics for RQ2 are also obtained from the results of literature review, See Section 3. After comparison of the results obtained, the model with the best performance is to be used in sales forecasting by the finance department of Volvo Trucks.

Chapter 2

Background

Sales forecasting is an important part of modern business intelligence. Forecasting of the future demand in sales is key to the planning and activity of trade and business. Forecasting helps business organizations to make improvements, to make changes in business plans and provides a solution related the stock storage. Speaking at an organizational level, forecasting plays a major role in decision activities in the areas wing to the essential sales, production, purchasing, finance, and accounting [9].

Sales is considered as a time series and sales forecasting is a major issue owing to the unpredictability of demand which relies upon numerous factors [10]. It can be a complex problem, especially in the case of a lack of data, missing data, and the presence of outliers. At present, different time series models [11] like Holt-Winters, ARIMA, etc, are in usage.

Sales prediction is preferably a regression problem than a time series problem. One of the main assumptions of regression methods is that the patterns in the past data will be repeated in future [10]. Implementations show that the use of regression approaches can often produce better results compared to time series methods [10]. By using supervised machine-learning methods like Random Forest [12], Gradient Boosting Machine [13], etc., It is possible to find complicated patterns in sales dynamics.

In this thesis, various machine learning regression techniques have been used to forecast the sales of the truck components and to find the best performed technique based on the results.

2.1 Time Series

Time series data consists of two obligatory components namely time units and the consonant value assigned for the provided time unit. Time series tracks the trends of the data points in a particular time period at regular intervals. In time series, there are no mandatory data size specifications that to be included, enabling the data to be congregated in such a way that it provides the information to be done by the analyst or person who examines the activity [14]. Time series is known to be as a widespread problem of significant practical interests as it allows to predict the future values of series with some margin of error from its past values [15].

Time series data is categorized into two types:

- **Stationary Time Series**: Stationary time series is one whose statistical properties such as the mean, variance and auto-correlation are all constant over time.
- **Non-Stationary Time Series**: Non-Stationary time series is the time series whose statistical properties such as the mean, variance and auto-correlation changes over time.

Non-stationary data is unpredictable and cannot be modeled or forecasted because of change in mean, variance and co-variance over the time. In order to achieve consistent, reliable results, it should be first converted into stationary data before performing further statistical analysis. For example, if the series is consistently increasing over time, the sample mean, and variance will grow with the size of the sample and they will always underestimate the mean and variance in future periods.

2.1.1 Univariate Time Series

A univariate time series is a series with a single time-dependent variable. In this type, only one variable will be varying over time. For example, the below sample consists of the hourly temperature values. Here, the temperature is the time dependent variable [16].

Time	Temperature
5:00 am	59 °F
6:00 am	60 °F
7:00 am	62 °F
8:00 am	59 °F
9:00 am	57 °F
10:00 am	58 °F
11:00 am	52 °F
12:00 pm	52 °F
1:00 pm	65 °F
2:00 pm	65 °F

Figure 2.1: Univariate Time Series

2.1.2 Multivariate Time Series

A Multivariate time series consists more than one time-dependent variable. Every variable depends not only on its past values but also has some dependency on other variables. For the same example in the univariate analysis, the below data has been included with more factors along with the temperature. In this case, there are multiple variables to be considered to optimally predict temperature. A series like this would fall under the category of multivariate time series. Figure 2.2 represents this:

Time	Temperature	Dew point	Humidity	Wind
5:00 am	59 °F	51 °F	74%	8 mph SSE
6:00 am	60 °F	51 °F	75%	8 mph SSE
7:00 am	62 °F	52 °F	75%	7 mph SSE
8:00 am	59 °F	52 °F	76%	7 mph S
9:00 am	58 °F	52 °F	76%	7 mph S
10:00 am	57 °F	54 °F	77%	8 mph SSW
11:00 am	52 °F	54 °F	77%	8 mph SSW

Figure 2.2: Multi-Variate Time Series

Multi-Variate time series forecasting can be done only on the availability of the feature variables historical data. It is possible to forecast the target variable using univariate time series data which is also called as univariate time series forecasting but in multi-variate time series analysis, it is not possible to forecast the target variable without the data of feature variables. This thesis uses multi-variate time series forecasting, where the historical sales data of feature variables also considered for forecasting the sales.

2.2 Machine Learning

Britannica Academic states, "Machine learning, in artificial intelligence (a subject within computer science), the discipline concerned with the implementation of computer software that can learn autonomously" [17]. There are a few sorts of issues that Machine learning tackles. Contingent upon the issue and data available, various approaches can be taken.

A supervised learning approach is when the data consists of both the input and output values. By calculating the error from the difference between the model's predicted value and the actual output value it is possible to change the model's weights and biases to minimize this error [18]. Supervised learning solves regression and classification problems. By using classification data can be categorized into a class, for example, red or car. By using regression, it is conceivable to estimate a numerical output value, for example, the number of cookies that are going to get sold.

Unsupervised learning is used when there is no output data. Instead, the algorithm tries to find patterns in the data on its own. Unsupervised learning is used to find clusters and put unseen data in a suitable neighborhood [18]. This can be used for example in sales patterns: if a customer buys milk hat he is likely to buy eggs.

With the input dataset used in this thesis and its numerical output value, it is clear that this is a regression problem. As there are more than one time-dependent variables in the chosen data, it is considered to be multi variate time series and the supervised learning method is preferred over an unsupervised learning method due to the availability of the data.

2.3 Machine Learning Algorithms

Sales prediction is preferably a regression problem than a time series problem, the utilization of the machine learning regression algorithms for sales forecasting can help in finding out better results when compared to the classical time series methods. Machine learning algorithms like Support Vector Regression, Random Forest Regression, Ridge Regression, and Gradient Boosting Regression can help to find put better results when compared with the traditional time series analysis methods.

2.3.1 Random Forest Regression

Random Forest (RF) had been proposed by Brieman [12], even though huge numbers of the ideas had proposed before in the writing in various literature. Breiman et al. [19] described CART (Classification and Regression Trees) was described as a non-parametric method for supervised learning who introduced and later introduced the bagging method to reduce for cart in 1996.

Random Forest Regression (RFR) is an ensemble method and a popular statistical learning method that extracts multiple samples from the original samples forecasting by using the bootstrapping method and combining the decision trees in order to perform them. RFR takes the mean of the predictions to get the results [20]. Random Forest can be said as a kind of additive model which makes predictions by making decisions combine from a sequence of base models [21]. This can be conventionally described as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots$$

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. While growing the trees, Random Forest adds additional randomness to the model [21].

Random Forest has nearly the same hyper parameters as a decision tree or a bagging classifier. Random Forest adds additional randomness to the mode while growing the trees [21]. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Random forest is popular for higher prediction speed and low memory usage because it chooses only a subset of features from the whole for splitting the node [21].

2.3.2 Support Vector Regression

Support vector machine (SVM) is a prominent machine learning model used for classification and regression purposes and it has been first introduced by Vapnik. SVM consists of two main categories: support vector classification (SVC) and support vector regression (SVR). SVM is a kind of learning system using a higher level dimensional feature space which yields prediction functions that are broadened on a

subset of support vectors [22]. Support Vector Regression is introduced in 1997 by Vapnik and two others [23].

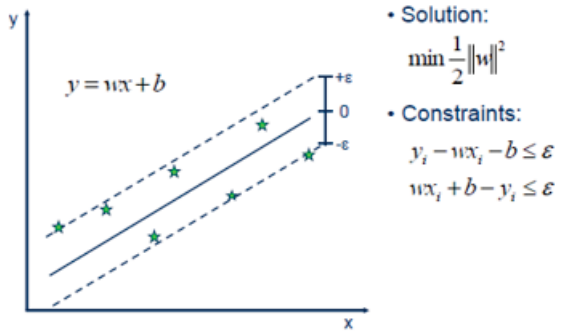


Figure 2.3: Support Vector Regression
[24]

In the above-represented figure, xi exhibit the predictors, yi exhibits dependent variable and ξ exhibits as the threshold where all the prediction values should be within the range.

2.3.3 Ridge Regression

Ridge regression is a method used for the analysis of the multiple regression data that suffer from multi-collinearity [25]. When the occurrence of the multi-collinearity happens, least squares are unbiased but their variances are large so they may be far from the true value [25]. By the addition of a degree of bias to the regression estimates, the standard error is reduced by the ridge regression. It is expected that the net effect will be to give estimates that are more reliable [25].

The cost function for ridge regression is as follows:

$$\min\left(\left|\left|Y - X(\theta)\right|\right|_2^2 + \lambda||\theta||_2^2\right)$$

Where X is a vector of weights, (θ) is the coefficient, (λ) is denoted by the alpha parameter in the ridge function. So, by altering, the values of the alpha penalty term are controlled. As the alpha value is higher, bigger is the penalty and accordingly coefficients magnitude is reduced. As the parameters shrink, it is mainly used for the of the multi-collinearity [26]. By doing the coefficient shrinkage model complexity is reduced and this process is called regularization.

2.3.4 Gradient Boosting Regression

Gradient Boosted Regression is considered to be one of the most effective machine learning models for predictive analytics [27]. Boosted trees model is a type of additive model that makes predictions by combining decisions from the base models [28].

Formally it can be written as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \dots$$

Where the final classifier g is the sum of simple base classifiers fi. For boosted trees model, each base classifier is a simple decision tree [27]. This broad technique of using multiple models to obtain better predictive performance is called model ensembling [27].

Unlike Random Forest which constructs all the base classifier independently, each using a subsample of data, GBRT uses a particular model ensembling technique called gradient boosting [27]. For a function f(x), assuming f is differentiable, gradient descent works by iteratively find

$$x_{t+1} = x_t - \eta \frac{\partial f}{\partial x}\bigg|_{x=x_t}$$

Where η is called the step size, x is a vector and the value of x found in the current iteration is its value in the previous iteration added to some fraction of the slope and gradient at this previous value.

Gradient Boosting Trees are very good at handling tabular data with numerical features, data consisting of categorical features with fewer than hundreds of categories. The boosted model can capture the non-linear interaction between the features and the target where liner models cannot do so [27].

2.3.5 Selection of Algorithms

To choose algorithms for an issue is not a trifling decision. There is definitely not an ideal algorithm that works for every problem, yet few algorithms are known to perform better on specific problems over others. In this case, algorithms like Support Vector Regression, Random Forest Regression, Gradient Boosting Machine, and Ridge Regression were expected to perform well on regression problems, given that it had incredible success in similar and comparable problems, see Section 3.

Chapter 3

Related Work

In this thesis, a literature review has been conducted for the finding suitable prediction model for forecasting the sales by looking at past research done in time series forecasting in the financial area using different machine learning algorithms.

Due to the importance of forecasting in many fields, many prominent approaches have been developed. In general, these methods termed as statistical methods, machine learning methods, hybrid methods. In time series analysis [29], Autoregressive (AR), Autoregressive Integrated Moving Average (ARIMA) [30] and Exponential smoothing methods [29, 31] are widely practised. Cox and Loomis [5] investigated the advancements in the field of forecasting of the last 25 years by inspecting the books related to the forecasting and thus the advancement of the forecast is noticed.

Francis E.H. Tay and Lijuan Cao [32] dealt with the application of a novel neural network technique in financial time series forecasting, support vector machine (SVM) to examine the feasibility of SVM in financial time series forecasting and proposed that SVMs achieve an optimum network structure by implement the structural risk minimization principle which seeks to minimize an upper bound of the generalization error rather than minimize the training error. This eventually results in better generalization performance than other neural network models. SVMs have also extended to solve non-linear regression estimation problems and they have been shown to exhibit excellent performance [23, 33].

Mariana Oliveira and Luis Torgo [34] made an attempt with the ensembles aiming for the improvement of prediction performance of the forecasting. and recognized ensembles as one of the most ambitious forms of solving predictive tasks and conventional in reducing the variance and bias components of the forecasting error by taking advantage of diversity amid models. Authors used the different sizes of embed with the summarization of the recently observed values. In this research study standard, bagging is compared along with the standard ARIMA and positive results are achieved showing that the approach is promising and can be used as an alternative for forecasting the time series. RMSE is used for performance evaluation [34].

Financial time series forecasting is inevitably a center point for the practitioner for its available data and for its profitability. Ensemble algorithms are substantial in improvising the performances of the base learners in financial time series forecasting. Authors experimented the Shanghai Composite Index dataset with the four

base learner algorithms, SVR (support vector regression), BPNN (back-propagation neural network), RBFNN (radial basis function neural network), locally weighted learning (LWL). Random Subspace, Stacking, and Bagging are used for comparison and evaluation of results show that the bagging provides stable improvement for the chosen dataset. RMSE and RAE are used as performance evaluation metrics [35].

In this paper, Authors [36] experimented financial time series forecasting by using an intelligent hybrid model to overcome the issue of capturing the non-stationarity property and to identify the accurate movements. In this study, empirical mode decomposition (EMD) along with support vector regression (SVR) is introduced and a novel hybrid intelligent prediction algorithm is proposed. EMD is used to decompose a non-linear and non-stationary signal into IMFs. In the experiment, proposed algorithm divides the data into intrinsic modes (IMF) by using EMD and moreover SVR with different kernels functions are used to the IMFs [36]. The outcome of the results shows that EMD-SVR model provides accurate results when compared to individual SVR results. RMSE, MAE, MAPE are used to evaluate the performance of models.

As financial time series forecasting, and modeling is quite difficult because of noise and non-stationarity presence. Authors [37] proposed a nonlinear radial basis function neural network ensemble model on dataset S and P index series and Nikkei 225 index series collected from the DataStream. The experiment has been carried out in four stages, primarily the data is divided into separate training sets by using bagging and boosting. In the next stage, training sets are given as input to individual RBF-NN models and relying on the diversity principle various predictors are produced. In the third stage, suitable neural network ensembles are chosen based on the Partial Least Squares technology (PLS). In the last stage, SVM-Regression is used for an ensemble of RBF-NN for forecasting. Experimental results show that the proposed ensemble method is better than some existing ensemble approach.

In this article, Authors [38] primary aim is to define factors that influence sowing crop sales quantities and create a technique for the most precise forecasting of their sales to support decision-making and enhance the effectiveness of agro-industrial companies company procedures. This paper also discussed an approach to the forecasting of sowing crop sales quantities, including the identification of variables affecting sales and the comparison of techniques for building mathematical models.Linear regression techniques, random forests and a neural network are used to construct forecasts. RMSE, ME and MAPE are used as evaluation metrics. Results have shown that Random Forest has produced better forecasts compared to neural networks [38].

From our research of previous studies on time series forecasting, it has been observed that the variety of machine learning models like Support Vector Machines, boosting methods, etc. have been used for regression problems. The accuracy values for forecasts are generally measured in RMSE, MAE. Table 3.1 is the Summarization of the literature review.

Motivation for the Research	**Results of the Research**
Investigated other methods beyond Neural Nets and ARIMA [39]	1. SVM models also succeeded in producing good forecasting results for retail demand. 2. RMSE is used to evaluate the performance
Discussed the study of standard bagging methods and compared results with the ARIMA [34]	1. Bagging methods produced promising results and stated that bagging can be used as an alternative for time series forecasting. 2. RMSE is used to evaluate the performance
Discussed the performances of the GBM methods for the forecasting [40]	1. GBM surpassed XGBoost, SGB in terms of computational speed and memory consumption. 2. RMSE and MAE are used to evaluate the performance
Investigated the performances of SVR and multi-layer back-propagation neural network [41]	1. Prediction results showed that performance of SVM is better than neural network. 2. RMSE and MAE are used to evaluate the performance
Investigated the ensemble methods with the time series data [42]	1. Ensemble methods have produced better results on financial time series data when compared with the ARIMA model. 2. RMSE and MAE are used to evaluate the performance.
Performed time series analysis on sales of sowing crops using machine learning methods [38]	1. Random Forest model outperformed Neural Networks. 2. RMSE, MSE are used as performance metrics
Investigated machine learning methods for forecasting the sales of retail stores [43]	1. Results shown that Ridge Regression outperformed Lasso Regression, Polynomial Regression and Linear Regression. 2. RMSE and R^2 are used as evaluation metrics.

Table 3.1: Summarization of the literature review

Chapter 4

Method

In this thesis, following research methods have been used to address research questions. First, in order to synthesize the results, a literature review is used to study the current existing relevant literature. These literature review results are used as an input to the analysis and assessment of the second research method experiment.

4.1 Literature Review

A **literature review** is conducted at the beginning of research to distinguish the various methodologies used to address various issues in sales forecasting. Benefits of each model and their principle for better execution for a specific issue is distinguished, four suitable regression techniques and two performance metrics have been selected based on literature review.

Inclusion Criteria:

- Articles which have been published between the years 2000 - 2019.
- Articles published in books,Journals, conferences and magazines .
- Articles that are in English language.
- Articles available in full text.

Exclusion Criteria:

- Articles not published in English.
- Articles without complete text.

4.2 Experiment

An **experiment** is chosen as a second research method in this thesis because the experiment is contemplated as the most suitable research method when dealing with the quantitative data and experiments give more control over factors and a greater subject size over other research techniques like a survey or a case study [16].

The main goal of this experiment is to evaluate the performance of machine learning models **Support Vector Machine Regressor, Ridge Regressor, Random Forest Regressor and Gradient Boosting Regressor** on the sales data extracted from the sales database of Volvo Trucks which is also the experimental data for this thesis. Results obtained from the experiment are analyzed and compared to select the best-performed algorithm among them for the chosen data. The independent and dependent variables of the experiment are as follows:

Independent Variables: Support Vector Machine Regressor, Ridge Regressor, Random Forest Regressor, Gradient Boosting Regressor, Size of the dataset.

Dependent Variables: Performance Metrics i.e. Mean Absolute Error and Root Mean Squared Error.

The procedure followed in this thesis can be contemplated as:

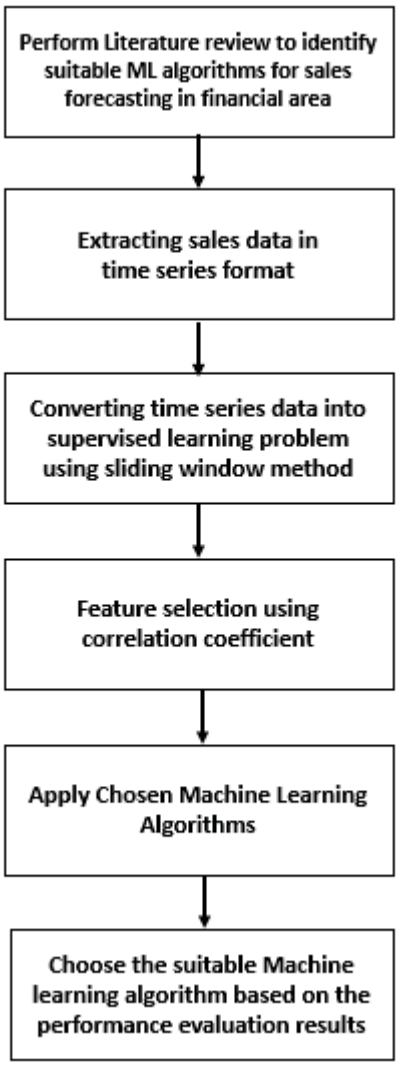


Figure 4.1: Mechanism followed in this thesis

4.3 Software Environment

4.3.1 Python

Python is a general-purpose, a high-level programming language designed to be easy to read and simple to implement. It is open source, even for commercial applications. Various programming features of python have been used for conducting an experiment in this thesis. The following are the libraries used:

Pandas

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. In specific, it offers data structures and operations for manipulating numerical tables and time series [44].

Numpy

Numpy is the fundamental package for scientific computing with Python and has powerful-dimensional array object which is used in the experiment along with the large collection of high-level mathematical functions to operate on these arrays [28].

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms [45].

Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics [46].

Sklearn

Sklearn is an open source library having efficient tools for data mining and data analysis. It provides the scope of supervised and unsupervised algorithms [26].

Machine Learning Algorithms used for experimentation in this thesis by using sklearn are:

- Random Forest Regressor
- Support Vector Regressor
- Gradient Boosting Regressor
- Ridge Regressor

Following are the performance evaluation metrics tools imported from sklearn,

- mean_squared_error
- mean_absolute_error

4.3.2 Jupyter Notebook

Jupyter Notebook is an open-source web application that allows to create and share documents that contain live code, equations, visualizations and narrative text [47]. In this thesis, Jupyter notebook is used to write the scripts for doing the experiment.

4.4 Dataset description

The dataset used in this thesis consists of the weekly sales of truck components from the year 2014 to 2019 extracted from the Volvo Trucks Sales database. It consists of 22 components of the truck including the target component belonged to the core component family. The final processed data consists of 260 weeks of data. The data is divided into training and testing data, 208 weeks of data is used for training models and 52 weeks data was unused and kept offset for testing. Time Series Split has been used for dividing the training and test set in a continuous manner which has been detailed in the next sections.

4.5 Data Preprocessing

Primarily, more than 12,000 truck components are recorded in the dataset, 22 components belonged to the family of core components have been used for the experiment. The dates where there was no sale for a specific truck component has been replaced with the zero. Moreover, sales parameter has been used as a feature in the models. Data of several components has been resampled from daily to weekly and then transformed into a supervised learning dataset. Sliding window method approach has been used for conversion because of its ease of implementation, low consumption of computation and memory.

4.5.1 Categorical Encoding

Categorical data are the variables that consist of the label values instead of numerical values. This type of variables is generally called as Nominal Variables. Some of the machine learning algorithms like decision trees don't require any numerical as it is able to learn from categorical data but most of the machine learning models require input and output variables to be numeric [48]. For the efficient implementation of machine learning algorithms Categorical data must be converted into a numerical data format. Following are the several encoding techniques for the transformation of data [48].

One Hot Encoding

It is one of the most prominently used approaches, this method compares each level of the categorical variable to a fixed reference level. For every category, a binary column is created and a dense array is returned. This can be said in a simple way as performing the binarization of category and including them as a feature [49].

As the data of some truck variants is categorical, one hot encoding has been used to represent categorical variables as binary vectors. Figure 4.2 represents the conversion of categorical data to binary vector.

Date	chassis height
2014-01-02	CHH-HIGH
2014-01-02	CHH-HIGH
2014-01-02	CHH-HIGH
2014-01-02	CHH-MED
2014-01-02	CHH-MED
2014-01-02	CHH-MED
2014-01-02	CHH-MED
2014-01-02	CHH-MED
2014-01-03	CHH-MED
2014-01-03	CHH-MED

(a) Categorical Data of Chassis Height

Date	chassis height_CHH-HIGH	chassis height_CHH-MED
2014-01-02	1	0
2014-01-02	1	0
2014-01-02	1	0
2014-01-02	0	1
2014-01-02	0	1
2014-01-02	0	1
2014-01-02	0	1
2014-01-02	0	1
2014-01-03	0	1
2014-01-03	0	1

(b) After applying one hot encoding

Date	chassis height_CHH-HIGH	chassis height_CHH-MED
2014-01-05	3.0	7.0
2014-01-12	0.0	17.0
2014-01-19	2.0	58.0
2014-01-26	4.0	39.0
2014-02-02	14.0	58.0
2014-02-09	2.0	42.0
2014-02-16	5.0	62.0
2014-02-23	5.0	100.0
2014-03-02	2.0	80.0
2014-03-09	4.0	70.0

(c) Re-sampling daily data to weekly

Figure 4.2: One hot encoding

4.5.2 Sliding window method

Sliding window method is used to transform the sequential supervised learning problem into the classical supervised learning problem [50].

```
1  time, Sales1, Sales2
2  1, 0.4, 66
3  2, 0.5, 69
4  3, 0.8, 87
5  4, 0.93, 98
6  5, 1.0, 99
```

Figure 4.3: Sample Time Series Data

By using the sliding window method, the time series problem in figure 4.3 can be restructured into a supervised learning problem. Figure 4.4 represents the same time series problem but after the application of the sliding window method [50].

```
1  A1, A2, B1, B2
2  ?, ?, 0.4 , 66
3  0.4, 66, 0.5, 69
4  0.5, 69, 0.8, 87
5  0.8, 87, 0.93, 98
6  0.93, 98, 1.0, 99
7  1.0, 99, ?, ?
```

Figure 4.4: After applying sliding window

Some of the comparison observations between the transformed dataset and the original time series:

- It can be seen in figure 4.4, the previous time step is the input (A1, A2) and the next time step is the output (B1, B2) in the supervised learning problem.
- The first and last row must be removed as the previous value will be used to predict the first value in the sequence.
- Order between the observations is preserved [50].

This issue in figure 4.4 can be contemplated in, where A1, A2, B1 can be the predictor factors and B2 can be a response variable. Here, the point to be noted is that the same problem can be composed in numerous different ways relying upon the necessities. Partial Auto Correlation Function (PACF) analysis is performed for determining the width of the sliding window or lag value [51]. However, before performing this method the time series data must be checked for stationarity. In this thesis Augmented Dickey-Fuller test is used for checking the stationarity.

4.5.3 Augmented Dickey-Fuller test

The Augmented Dickey-Fuller test is a type of statistical test that tests the null hypothesis that a unit root is present in an autoregressive model. The alternative

hypothesis depends on which version of the test is used but is usually stationarity or trend-stationarity [52].

The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary (has some time-dependent structure) [52]. The alternate hypothesis is that the time series is Stationary [52].

- **Null Hypothesis (H0)**: If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time-dependent structure.
- **Alternate Hypothesis (H1)**: The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have a time-dependent structure [53].

The result can be explained as follows:

- **P-value $>$ 0.05:** Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
- **P-value $<=$ 0.05:** Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

4.5.4 Normalization

Normalization is the process of rescaling the data from original range to a specific scale [53]. Normalization is used to transform the features to a fixed range to eliminate the influence of one feature over the other. In this thesis, Min-Max Normalization is performed using the scikit-learn object MinMaxScalar by implementing the following formula.

$$\hat{x_n} = \frac{x_n - x_n(min)}{x_n(max) - x_n(min)}$$

Where min and max are the minimum and maximum values in x given its range.

Date	adr adaption_ADR2	adr adaption_UADR	adr classification_ADRC-FL
2014-01-05	0.0	10.0	0.0
2014-01-12	2.0	15.0	2.0
2014-01-19	4.0	54.0	4.0
2014-01-26	4.0	39.0	4.0
2014-02-02	10.0	62.0	10.0
2014-02-09	11.0	33.0	11.0

(a) Before Normalization

Date	adr adaption_ADR2	adr adaption_UADR	adr classification_ADRC-FL
2014-01-05	0.000000	0.256410	0.000000
2014-01-12	0.051282	0.384615	0.051282
2014-01-19	0.102564	1.384615	0.102564
2014-01-26	0.102564	1.000000	0.102564
2014-02-02	0.256410	1.589744	0.256410
2014-02-09	0.282051	0.846154	0.282051

(b) After Normalization

4.6 Feature Selection

There are many factors which might affect the success of machine learning model on a given task. Feature selection is one of the core concepts which will have a huge influence on the model performance. The data features which are selected for training the machine learning model will have a big impact on the performance because irrelevant features negatively influence the performance of the model. Implementation of the feature selection process helps in upgrading the performance of the predictor, reducing overfitting by removing the data redundancy, reduces the training time and improves the accuracy of the model. In this thesis, following approach have been used for the feature selection.

4.6.1 Correlation method

Correlation states how the features are related or can be an as mutual statistical dependency between each other or with the chosen target variable. Correlation can be negative (if an increase in one value of feature decreases the value of the other feature) and positive (if the increase in the one value of the feature increases the other values of the features). If there's no impact between the values of the features then it can be said as no correlation exists [54]. There are different methods for the calculation of correlation coefficient, Pearson correlation coefficient measures the linear association between the continuous variables and the Spearman correlation coefficient measures the relationship between the variables by using the monotonic function.

Spearman correlation coefficient is a popular method for dealing with the problems related to the non-linear data, so it is used for dealing with the multivariate data in finding out the associations between the variables. In this thesis, as most part of the data is non-linear, Spearman correlation coefficient has been preferred [55].

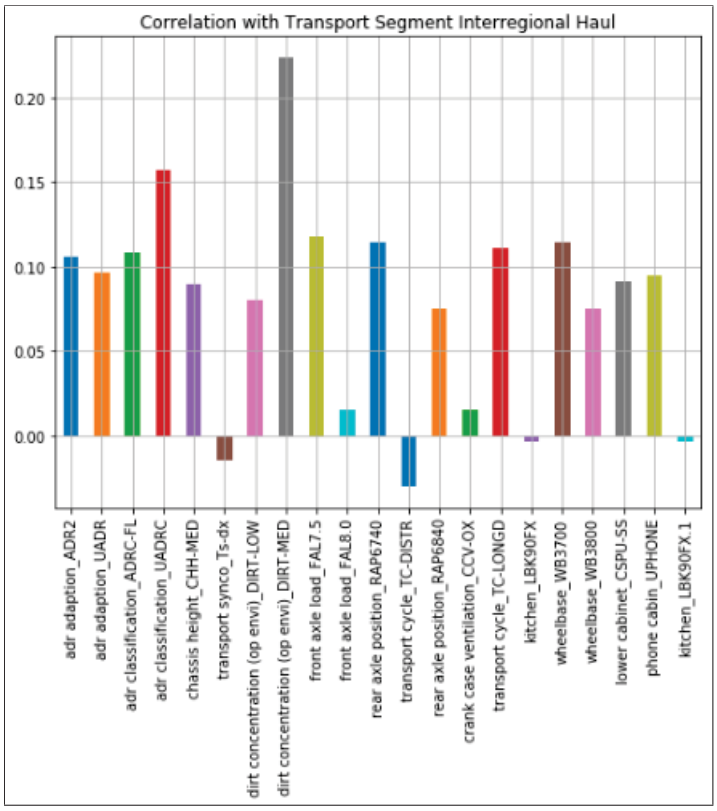


Figure 4.6: Correlation with the target Variant

Therefore, in this thesis, Spearman Correlation coefficient is used for feature selection. Variables which have a negative correlation with the target variable had been removed to improvise the performance of the machine learning model, figure 4.6 represents the correlation with the feature variants and target variant.

4.7 Experimental Setup

- Performed 5-fold time series cross-validation with Support Vector Machine Regressor, Ridge Regressor, Random Forest Regressor and Gradient Boosting on the data-set.
- Performances of the algorithms have been experimented and the results are compared for selecting the best-performed algorithm for this dataset.

4.7.1 Performance Metrics

Determination of the performance evaluation metrics ought to be made in compliance with the regression problem and the experimental data. Performance metrics chosen in this thesis are commonly used for the evaluation of the time series forecasting problems and also in the common machine learning problems generally used for regression. This additionally makes this thesis comparable to the other existing research works which are currently using the statistical techniques for forecasting. In the related work section, usage of the performance metrics by various authors have

been mentioned. Performance evaluation metrics used in this thesis are:

Mean Absolute Error:

Mean Absolute Error is a standard measure of forecast error in the time series analysis. MAE is one of the many metrics for summarizing and evaluating the performance of the machine learning model. The mean absolute is a quantity used to measure how close forecasts or prediction are to the eventual outcomes. As the name suggests, the mean absolute error is an average of the absolute errors [56]. Lower the error implies greater the accuracy of the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where y_i represents actual values and $\hat{y}_i$ represents the forecasted values.

Root Mean Square Error

Root Mean Square Error (RMSE) is the square root of the mean square error. It is the root of the average of squared differences between prediction and observation. Lower the error implies greater the accuracy of the model [56].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where y_i represents actual values and $\hat{y}_i$ represents the forecasted values.

Why RMSE is preferred over MAE?

Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. RMSE does not necessarily increase with the variance of the errors. RMSE increases with the variance of the frequency distribution of error magnitudes [57]. The Root Mean Square Error is minimized by the conditional mean whereas the MAE is by the conditional median. So, if the MAE is minimized then fit will be approximately closer to the median and will be biased.

4.8 Time Series Cross-Validation

Time Series Cross Validation is an approach consists of a series of test sets, each composed of a single observation. The corresponding test set consists of only observations that occurred prior to the observation that forms the test set [58]. The accuracy of the forecast is computed by averaging over the test sets and this procedure is sometimes known as "evaluation on a rolling forecasting origin" [58].

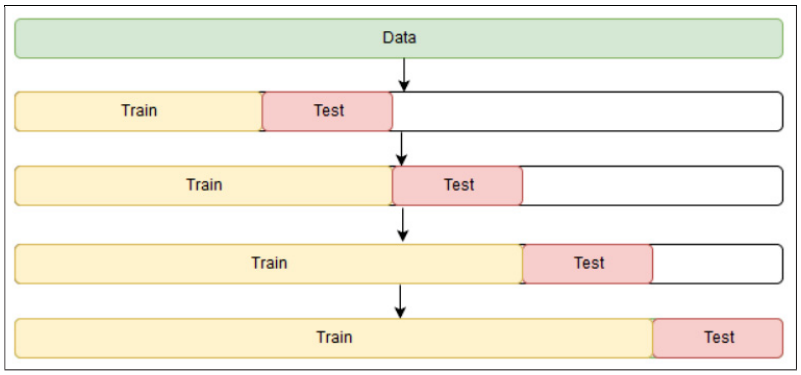


Figure 4.7: Time Series Cross Validation

Cross-validation for time series is used because time series consists of the temporal structure and values cannot be randomly mixed in a fold while preserving this structure. With randomization, all-time dependencies between observations will be lost. To overcome this issue, more deceptive approach has to be used where starting with a small subset of data for training purpose, making forecast for the later data points and then checking the accuracy for the forecasted data points. These data points are then included as part of the next training dataset and subsequent data points are forecasted. The error obtained on each split is averaged in order to compute a robust estimate of model error [59].

In simpler words, the model has to be trained on a small segment of time series from the beginning until sometime **t**, predictions for the next **t**+**n** steps has to be computed, and error needed to be calculated. Then the training sample to be extended to **t**+**n** value, predictions from **t**+**n** until **t**+**2n** steps has to be computed, and same procedure of moving the test segment of the time series is continued until the last available information. As a result, there will be as may folds as n will fit between the initial training sample and the last observation.

Chapter 5

Results

5.1 Stationarity Check

Stationarity of the dataset has been examined by using the augmented dicky fuller test. The below figure represents the results obtained:

```
Results of Dickey Fuller Test:
Test Statistic                  -5.316866
p-value                          0.000005
#Lags Used                       7.000000
Number of Observations Used     43.000000
Critical Value (1%)             -3.592504
Critical Value (5%)             -2.931550
Critical Value (10%)            -2.604066
dtype: float64
```

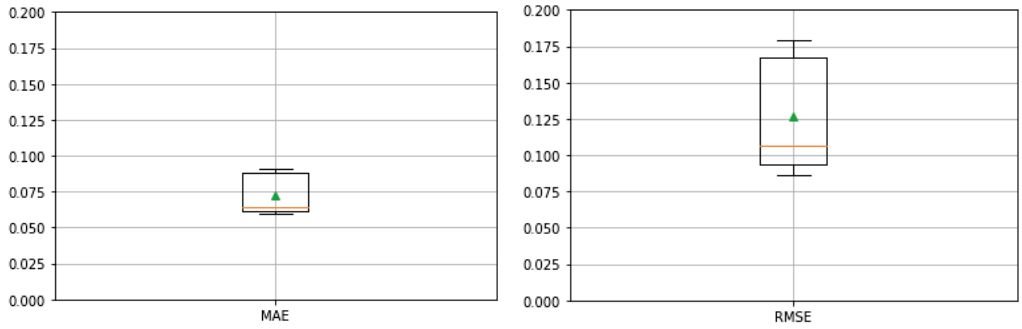
Figure 5.1: Augmented Dickey-Fuller Test

From the figure 5.1, it can be noted that the p-value is less than 0.5, hence the data is stationary and can be further proceeded for time series analysis. From the PACF analysis, window width for sliding window is chosen as 1 and this lag value is applied to all the feature variables in the data.

5.2 Gradient Boosting Regressor

Gradient Boosting Regressor is trained with the dataset by using 5-fold time series cross-validation approach where 80% of the data was used for training and 20% of the data was used as the test set and the performances have been measured by using the metrics, MAE, and RMSE. The following are the results obtained :

In Figure 5.2a, the box-and-whisker plot represents the Mean Absolute Error (MAE) acquired by Gradient Boosting Regressor on a 5-fold time series cross-validation tests. The upper whisker of the box-plot represents the maximum MAE which is **0.09068**, the middle quartile represents median MAE which is **0.064439** and the lower quartile of the box-and-whisker plot represents minimum MAE which is **0.0598**. The triangle in the box-and-whisker plot represents the mean MAE which is **0.072767**.



(a) MAE Box Plot (b) RMSE Box Plot

Figure 5.2: Gradient Boosting Error Box Plots

In Figure 5.2b, the box-and-whisker plot represents the Root Mean Squared Error (RMSE) acquired by Gradient Boosting Regressor on a 5-fold time series cross-validation tests. The upper whisker of the box-plot represents the maximum RMSE which is **0.17933**, the middle quartile represents median RMSE which is **0.10626** and the lower quartile of the box-and-whisker plot represents minimum RMSE which is **0.08634**. The triangle in the box-and-whisker plot represents the mean RMSE which is **0.12661**.

Figure 5.3 represents the actual and forecasted sales of target variant obtained using the Gradient boosting regression model where the orange line represents the actual values and blue line represents forecasted sales of the target variant interregional haul. This color notation is continued over the **Figures 5.5,5.7,5.9**.

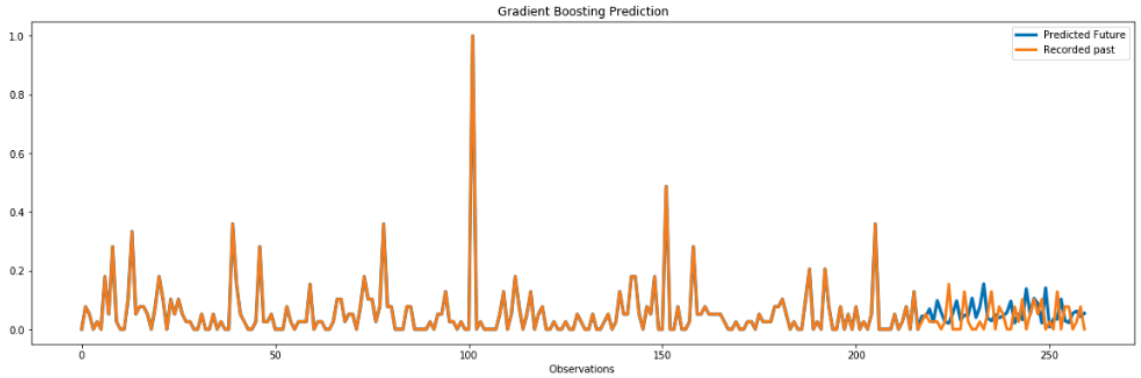
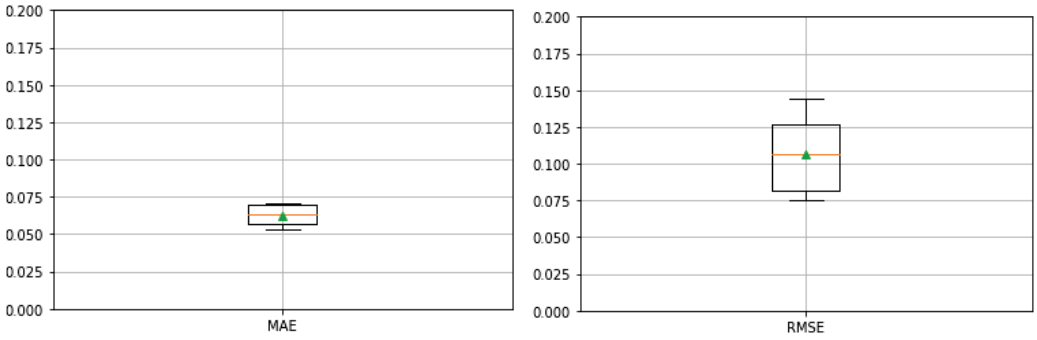


Figure 5.3: Gradient Boosting Regression Prediction

5.3 Random Forest Regressor

Random Forest Regressor is trained with the dataset by using 5-fold time series cross-validation approach where 80% of the data was used for training and 20% of the data was used as the test set and the performances have been measured by using the metrics MAE and RMSE.The following are the results obtained by the Random Forest Regressor:



(a) MAE Box Plot (b) RMSE Box Plot

Figure 5.4: Random Forest Regression Error Box Plots

In Figure 5.4a, the box-and-whisker plot represents the Mean Absolute Error (MAE) acquired by Random Forest Regressor on a 5-fold time series cross-validation tests. The upper whisker of the box-plot represents the maximum MAE which is **0.07096**, the middle quartile represents median MAE which is **0.06315** and the lower quartile of the box-and-whisker plot represents minimum MAE which is **0.05283**. The triangle in the box-and-whisker plot represents the mean MAE which is **0.062706**.

In Figure 5.4b, the box-and-whisker plot represents the Mean Squared Error (RMSE) acquired by Random Forest Regressor on a 5-fold time series cross-validation tests. The upper whisker of the box-plot represents the maximum RMSE which is **0.1441**, the middle quartile represents median RMSE which is **0.1064** and the lower quartile of the box-and-whisker plot represents minimum RMSE which is **0.07535**. The triangle in the box-and-whisker plot represents the mean RMSE which is **0.10683**. Figure 5.5 represents the actual and forecasted sales of target variant transport segment-interregional haul obtained using the random forest regression model.

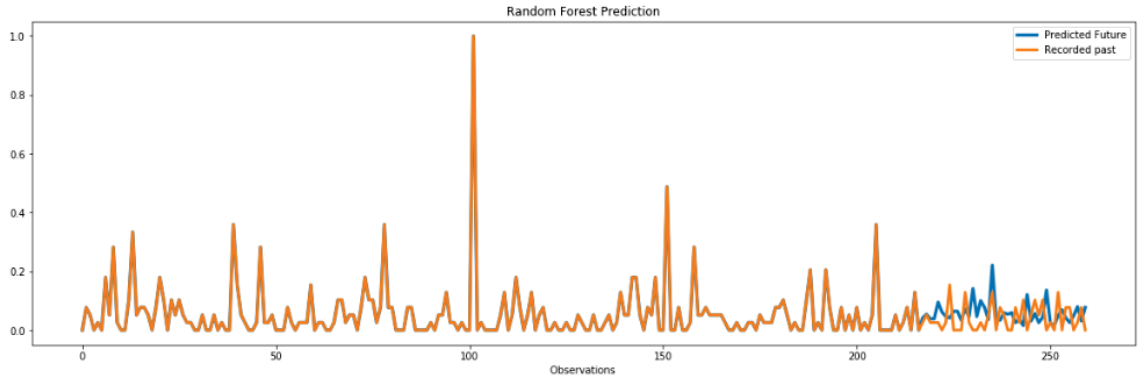
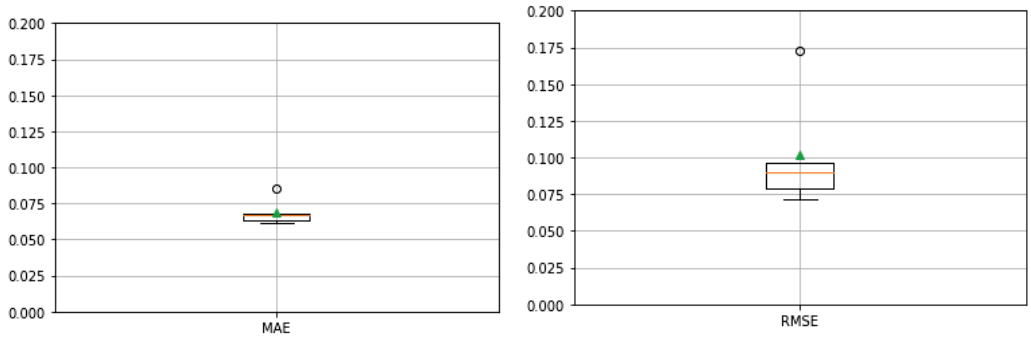


Figure 5.5: Random Forest Regressor Prediction

5.4 Support Vector Regressor (SVR)

Support Vector Regressor is trained with the dataset by using 5-fold time series cross-validation approach where 80% of the data was used for training and 20% of

the data was used as the test set and the performances have been measured by using the metrics MAE and RMSE. The following are the results obtained by the Support Vector Regressor:



(a) MAE Box Plot (b) RMSE Box Plot

Figure 5.6: Support Vector Regression Error Box Plots

In figure 5.6a, the box-and-whisker plot represents the Mean Absolute Error (MAE) acquired by Support Vector Regressor on a 5-fold time series cross-validation tests. The upper whisker of the box-plot represents the maximum MAE which is **0.08495**, the middle quartile represents median MAE which is **0.06697** and the lower quartile of the box-and-whisker plot represents minimum MAE which is **0.06102**. The triangle in the box-and-whisker plot represents the mean MAE which is **0.68752**

In Figure 5.6b, the box-and-whisker plot represents the Root Mean Squared Error(RMSE) acquired by Support Vector Regressor on a 5-fold time series cross-validation tests. The upper whisker of the box-plot represents the maximum RMSE which is **0.17254**, the middle quartile represents median RMSE which is **0.09035** and the lower quartile of the box-and-whisker plot represents minimum RMSE which is **0.07127**. The triangle in box-and-whisker plot represents the mean RMSE which is **0.10188**, figure 5.7 represents the actual and forecasted sales obtained using the Support Vector regression model

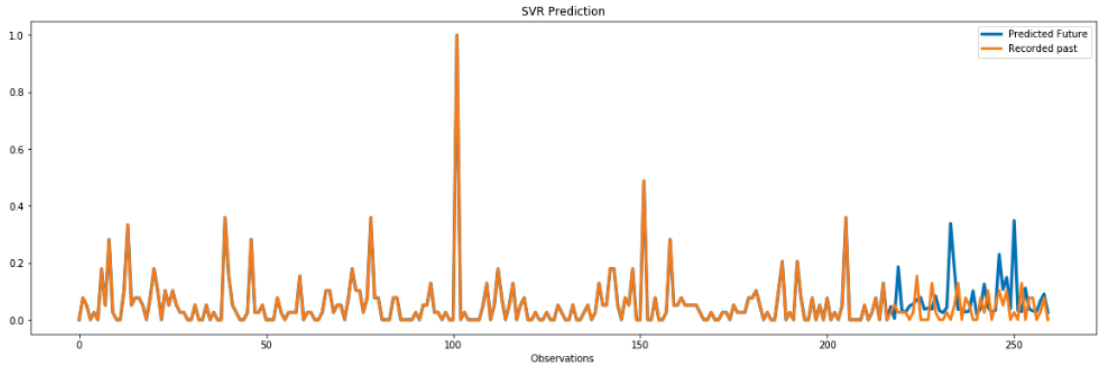
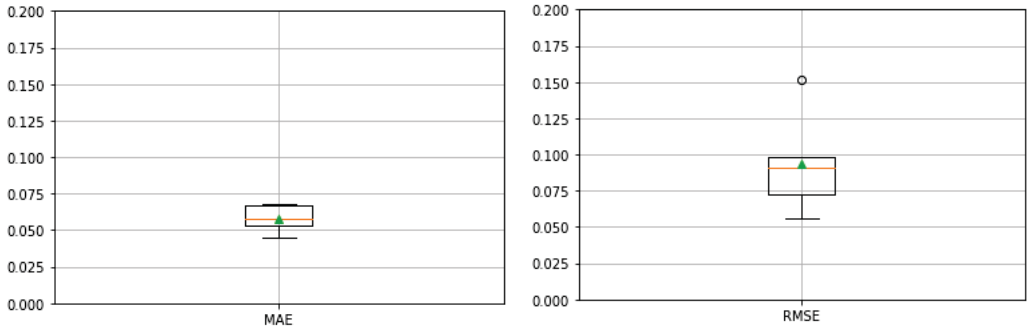


Figure 5.7: Support Vector Regressor Prediction

5.5 Ridge Regressor

Ridge Regressor is trained with the dataset by using 5-fold time series cross-validation approach where 80% of the data was used for training and 20% of the data was used as the test set and the performances have been measured by using the metrics MAE and RMSE. The following are the results obtained by the Ridge Regressor:



(a) MAE Box Plot (b) RMSE Box Plot

Figure 5.8: Ridge Regression Error Box Plots

In figure 5.8a, the box-and-whisker plot represents the Mean Absolute Error (MAE) acquired by Ridge Regressor on a 5-fold time series cross-validation tests. The upper whisker of the box-plot represents the maximum MAE which is **0.06814** the middle quartile represents median MAE which is **0.05779** and the lower quartile of the box-and-whisker plot represents minimum MAE which is **0.04508**. The triangle in box-and-whisker plot represents the mean MAE which is **0.058168**.

Figure 5.8b, the box-and-whisker plot represents the RMSE acquired by Ridge Regressor on a 5-fold time series cross-validation tests. The upper whisker of the box-plot represents the maximum RMSE which is **0.1512**, the middle quartile represents median RMSE which is **0.09114** and the lower quartile of the box-and-whisker plot represents minimum RMSE which is **0. 05572**. The triangle in the box-and-whisker plot represents the mean RMSE which is **0.09367**, figure 5.9 represents the actual and forecasted sales of target variant obtained using the ridge regression model.

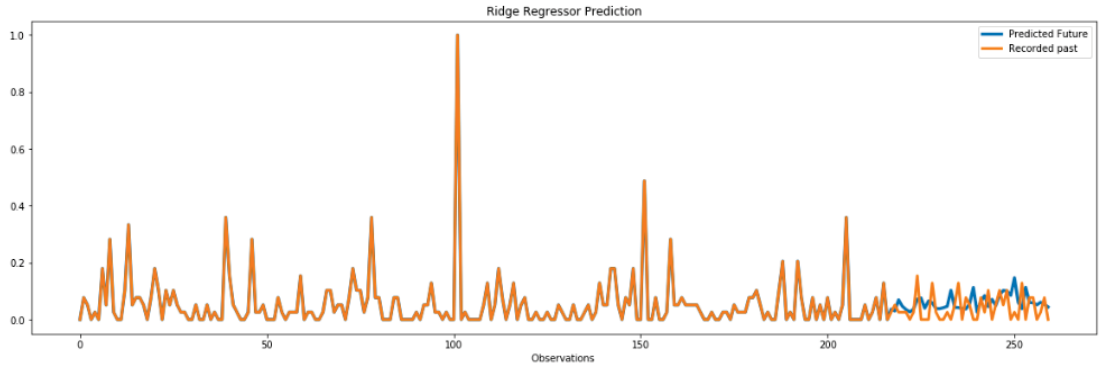


Figure 5.9: Ridge Regressor Prediction

5.6 Performance evaluation results

Algorithms	Mean Absolute Error	Root Mean Square Error
Ridge Regression	0.05816	0.09367
Support Vector Machine	0.06875	0.10188
Random Forest Regression	0.06270	0.10683
Gradient Bossting Regressor	0.07276	0.12661

Table 5.1: Comparison of performance evaluation results

From the table 5.1, Ridge Regression performed well with both the metrics MAE and RMSE, Ridge Regression has least error in forecasting the sales of truck components when compared to the Support Vector Machine, Gradient Boosting regression and Random Forest. The gradient boosting machine demonstrated the worst performance with the highest error in both metrics.

Chapter 6

Analysis and Discussion

6.1 Comparative study of Performance Metrics

6.1.1 Mean Absolute Error

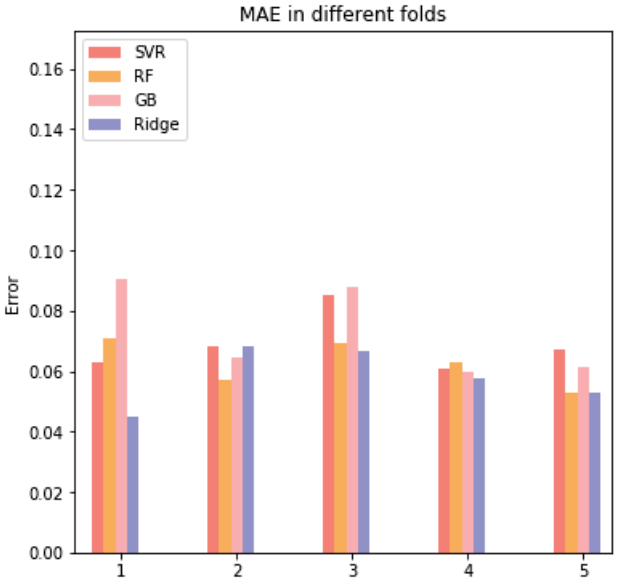


Figure 6.1: Comparison of MAE obtained by regression models on 5-fold time series validation tests

Figure 6.1 represents the Mean Absolute error from the results of the predictions produced by the Support Vector Regressor, Random Forest Regressor, Gradient Boosting Regressor and Ridge Regressor on 5-fold time series cross-validation tests. From the figure, it can be seen that except in 2nd fold, Ridge regressor has least MAE in all the folds and thus can be said as a best-performed algorithm. Gradient boosting has highest MAE and thus it can be said as worst performer.

6.1.2 Root Mean Squared Error

From figure 6.2, it can be noticed that Ridge Regression has shown spectacular overall performance with the least RMSE and Gradient Boosting Regressor has shown worst performance with the highest RMSE.

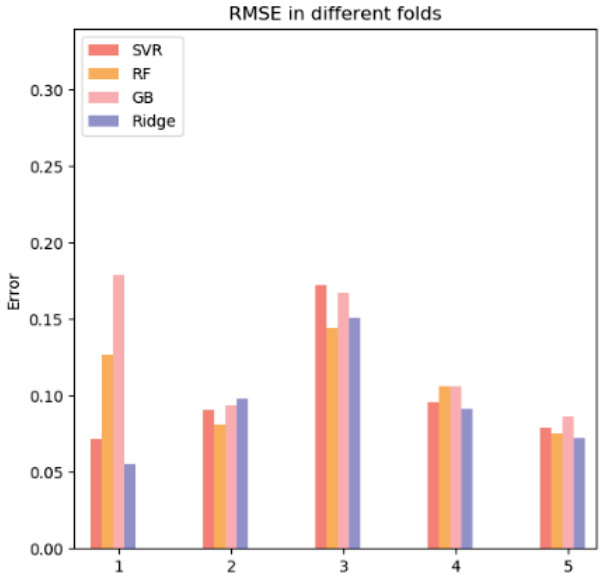


Figure 6.2: Comparison of RMSE obtained by regression models on 5-fold time series validation tests

6.1.3 Average Mean Absolute Error

From figure 6.3, The average MAE acquired by the Ridge Regression across the 5-fold time series cross-validation is 0.058168, followed by the Random Forest Regressor which is 0.062706, thereafter SVR which is 0.068752 and finally Gradient Boosting Regressor which is 0.072767. From figure 6.3, it can said that Ridge Regression is the best performer with the least error and Gradient boosting is the worst performer with the highest error.

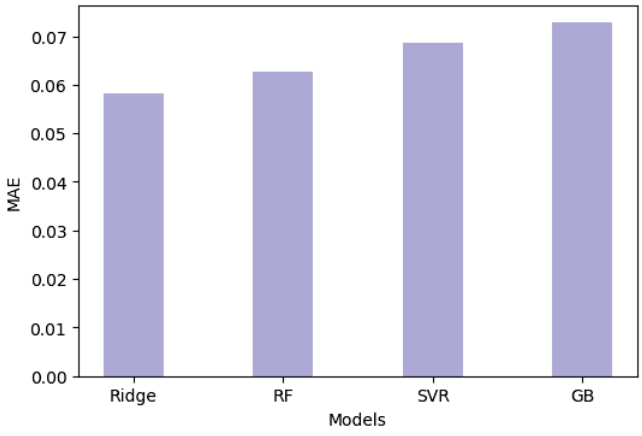


Figure 6.3: Comparison of Average MAE obtained by regression models on 5-fold time series validation tests

6.1.4 Average Root Mean Square Error

From figure 6.4, The average RMSE acquired by the Ridge Regression across the 5-fold time series cross-validation is 0.09367, followed by the SVR which is 0.101886, thereafter Random Forest Regressor which is 0.106836 and finally Gradient Boosting Regressor which is 0.12661. From figure 6.4, it can said that Ridge Regression is the best performer with the least error and Gradient boosting is the worst performer with the highest error.

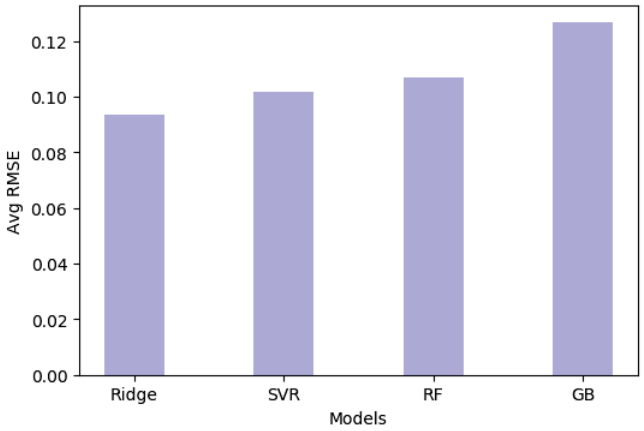


Figure 6.4: Comparison of Average RMSE obtained by regression models on 5-fold time series validation tests

6.2 Key Analysis

Key concepts behind the performance of algorithms are explored as follows:

- Ridge Regression has performed significantly well compared to the other models. This may be because of the regularization approach where the variance is reduced at the cost of some bias initiation which makes it robust to outliers and overfitting.
- Support Vector Regression has performed surprisingly well compared to the random forest and gradient boosting, this may be because of its generalization capability. Kernel function for the chosen parameters is set to 'rbf', other kernel functions like linear, radial may produce better results.
- Random Forest Regressor hasn't shown good performance, this may be because of overfitting problem which may be preventing from generalizing the model.
- Gradient boosting produced bad results compared to others because of overfitting problem and Gradient Boosting Regressor is harder to tune compared to Random Forest.

6.3 Discussion

RQ1: What are the suitable machine learning techniques for forecasting sales in the area of finance?

Answer: Based on the results obtained from the Literature review, four machine learning models namely Ridge Regressor, Support Vector Regressor, Random Forest Regressor, and Gradient boosting Regressor have been chosen for forecasting the sales.

RQ2: What are the performances of machine learning models for forecasting the sales of the core components of trucks based on the historical data?

Answer: Ridge Regression is the best suitable algorithm to forecast the sales of truck components.In this experiment, Ridge Regression has least error in forecasting the sales when compared to the Support Vector Machine, Random Forest, Gradient Boosting regression. This is because of the regularization, where slack variables are added to avoid over-fitting, Average RMSE across the 5-fold time series cross-validation is **0.09367** which is quite outstanding and Average MAE is **0.05779**. Gradient Boosting has shown worst performance compared to the other algorithms, Average RMSE across the 5-fold time series cross-validation is **0.12661** which is worst and Average MAE is **0.07276**. The performances of all the models have been discussed in **Section 6.1**

6.4 Contributions

Although there have been existing researches focusing on the sales forecasting by using the statistical techniques in automotive industry, there has been no research based on developing the sales forecasting related to the truck components based on the machine learning, which can be used for the sales team of any wide range of companies and organizations in the automotive industry. This thesis suggests that the machine learning approach can be used to get insights from the data and to forecast the sales of truck components. This approach can be further used to develop advanced forecasting tool which can able to forecast more precise forecasts.

6.5 Threats to Validity

The concept of validity was formulated by Kelly, who expressed that a test is valid if it measures what it claims to measure.

6.5.1 Internal Validity

Internal validity refers to the extent to which research was carried out [60], To overcome the threat of missing observations in the experiments, cloud backup is used which consists of all the logs copies of the experiment

6.5.2 External Validity

External validity is the validity of applying the conclusions of a scientific study outside the context of that study [60]. This validity is attained by the data extracted from the sales database which has been used in this thesis study to evaluate the algorithm and its performance. The risk of the particularity of variables is mitigated by describing all the dependent variables of this study in such a way that they are significant in any general experimental settings.

6.5.3 Conclusion Validity

Conclusion validity refers to if the data used from the experiment and results are justified and right [60]. This issue may be raised if there is no proper selection of performance evaluation metrics can lead to an understanding of the size of the relationship between independent and dependent variables in the study. To avoid this threat multiple evaluation metrics have been used along with the proper structure of experimental setup and methodology.

6.6 Limitations

- The study has been conducted on sales data which belonged to the truck sales of Sweden region and it cannot be assured that the similar results will be obtained from the study conducted on the sales data belonged to the other region as the sales may vary in other regions.
- Due to the unavailability of the company's information like customer details, certain campaigns and discounts. They haven't been included in data which would benefit in obtaining better forecasts.

Chapter 7

Conclusions and Future Work

Sales forecasting is a pivotal part of the financial planning of business for any organization. It can be said as a self-assessment tool which uses the statistics of the past and the current sales in order to predict future performance. Sales forecasting plays an important role in optimising the truck sales process. Financial and Sales planning with the help of the sales forecasts helps to get the information needed to predict the revenue as well as the profit. Thus, in finding such solution for sales forecasts, machine learning algorithms such as Random Forest Regressor, Support Vector Regressor, Ridge Regressor, and Gradient Boosting Regressor have been evaluated on Volvo truck components sales data which can forecast the short term sales and help the organization in making the key decisions. After performing the various statistical tests and performance metrics, it is found that Ridge Regression is a suitable algorithm in accordance to the chosen dataset and thus accomplishing the aim of this thesis.

7.1 Future Work

Future work for this thesis involves comparing the performance evaluation results of the chosen regression techniques obtained from this thesis with the results obtained from deep learning methods which could help the researchers in getting better trends and results.

As mentioned earlier, due to the lack of information about external or environmental variables like promotions, discounts, etc. Such variables are not considered in the modeling and sales forecasting experiment. It is suggested to use the related data of the company for further research to obtain more accuracy.

Bibliography

[1] Chris Chatfield. *Time-Series Forecasting*. en. Chapman and Hall/CRC, Oct. 2000. ISBN: 978-0-429-12635-2. DOI: 10.1201/9781420036206. URL: https://www.taylorfrancis.com/books/9780429126352 (visited on 05/21/2019).

[2] Chi-Jie Lu, Tian-Shyug Lee, and Chia-Mei Lian. "Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks". In: *Decision Support Systems* 54.1 (2012), pp. 584–596.

[3] James J Pao and Danielle S Sullivan. "Time Series Sales Forecasting". In: *Final Year Project* (2017).

[4] Mohit Gurnani et al. "Forecasting of sales by using fusion of machine learning techniques". In: *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*. IEEE, 2017, pp. 93–101.

[5] James E. Cox Jr and David G. Loomis. "Improving forecasting through textbooks—A 25 year review". In: *International Journal of Forecasting* 22.3 (2006), pp. 617–624.

[6] Michael Chui. "Artificial intelligence the next digital frontier?" In: *McKinsey and Company Global Institute* 47 (2017).

[7] Bruce Saunders. "A land development financial model". In: (1990).

[8] Christoph Freudenthaler, Lars Schmidt-Thieme, and Steffen Rendle. "Bayesian factorization machines". In: (2011).

[9] A. Syntetos. "John T. Mentzer and Mark A. Moon, Sales forecasting management: A demand management approach , Sage Publications, Thousand Oaks, London (2005) ISBN 1-4129-0571-0 Softcover, 347 pages". In: *International Journal of Forecasting* 22.4 (2006), pp. 821–821.

[10] Bohdan M. Pavlyshenko. "Machine-Learning Models for Sales Time Series Forecasting". In: *Data* 4.1 (2019), p. 15.

[11] Chris Chatfield and Time-Series Forecasting. "Chapman & Hall". In: *CRC, London/Boca Raton* (2001).

[12] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[13] Jerome H. Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[14] Jose Luis Blanco et al. "Artificial intelligence: Construction technology's next frontier". In: *Building Economist, The* September 2018 (2018), p. 8.

[15] Ahmed Tealab. "Time Series Forecasting using Artificial Neural Networks Methodologies: A Systematic Review". In: *Future Computing and Informatics Journal* (2018).

[16] Jason Brownlee. *Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future*. Jason Brownlee, 2017.

[17] N. Alan Heckert. *Statistical Engineering Division*. en. Aug. 2010. URL: `https://www.nist.gov/itl/sed` (visited on 05/21/2019).

[18] Andries P. Engelbrecht. *Computational intelligence: an introduction*. John Wiley & Sons, 2007.

[19] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.

[20] Juan Huo, Tingting Shi, and Jing Chang. "Comparison of Random Forest and SVM for electrical short-term load forecast with different data sources". In: *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2016, pp. 1077–1080.

[21] Turi. *Random Forest Regression | Turi Machine Learning Platform User Guide*. URL: `https://turi.com/learn/userguide/supervised-learning/random_forest_regression.html` (visited on 05/21/2019).

[22] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. "Support vector regression". In: *Neural Information Processing-Letters and Reviews* 11.10 (2007), pp. 203–224.

[23] Vladimir Vapnik, Steven E. Golowich, and Alex J. Smola. "Support vector method for function approximation, regression estimation and signal processing". In: *Advances in neural information processing systems*. 1997, pp. 281–287.

[24] Saed Sayad. *Real time data mining*. Self-Help Publishers Cambridge, 2011.

[25] J. Al-Jararha. "New approaches for choosing the ridge parameters". In: *Hacettepe Journal of Mathematics and Statistics* 47.6 (2016), pp. 1625–1633.

[26] *scikit-learn: machine learning in Python — scikit-learn 0.21.1 documentation*. URL: `https://scikit-learn.org/stable/` (visited on 05/21/2019).

[27] *Boosted Trees Regression | Turi Machine Learning Platform User Guide*. URL: `https://turi.com/learn/userguide/supervised-learning/boosted_trees_regression.html` (visited on 05/21/2019).

[28] *NumPy — NumPy*. URL: `http://www.numpy.org/` (visited on 05/21/2019).

[29] GE Box et al. *Time series analysis, control, and forecasting . Hoboken*. 2015.

[30] Mohit Gurnani et al. "Forecasting of sales by using fusion of machine learning techniques". In: *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*. IEEE. 2017, pp. 93–101.

[31] Peter R Winters. "Forecasting sales by exponentially weighted moving averages". In: *Management science* 6.3 (1960), pp. 324–342.

[32] Francis EH Tay and Lijuan Cao. "Application of support vector machines in financial time series forecasting". In: *omega* 29.4 (2001), pp. 309–317.

[33] K-R Müller et al. "Predicting time series with support vector machines". In: *International Conference on Artificial Neural Networks*. Springer. 1997, pp. 999–1004.

[34] Mariana Rafaela Oliveira and Luis Torgo. "Ensembles for time series forecasting". In: (2014).

[35] Cheng Cheng, Wei Xu, and Jiajia Wang. "A comparison of ensemble methods in financial market prediction". In: *2012 Fifth International Joint Conference on Computational Sciences and Optimization*. IEEE, 2012, pp. 755–759.

[36] L. I. N. Feng-Jenq. "Adding EMD Process and Filtering Analysis to Enhance Performances of ARIMA Model When Time Series Is Measurement Data". In: *Romanian Journal of Economic Forecasting* 18.2 (2015), p. 92.

[37] Donglin Wang and Yajie Li. "A novel nonlinear RBF neural network ensemble model for financial time series forecasting". In: *Third International Workshop on Advanced Computational Intelligence*. IEEE, 2010, pp. 86–90.

[38] Mohammed A Al-Gunaid et al. "Time series analysis sales of sowing crops based on machine learning methods". In: *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE. 2018, pp. 1–6.

[39] Andreas Zell. *Simulation neuronaler netze*. Vol. 1. 5.3. Addison-Wesley Bonn, 1994.

[40] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3146–3154.

[41] R Samsudin, A Shabri, and P Saad. "A comparison of time series forecasting using support vector machine and artificial neural network model". In: *Journal of applied sciences* 10.11 (2010), pp. 950–958.

[42] Yaohui Bai et al. "Forecasting financial time series with ensemble learning". In: *2010 International Symposium on Intelligent Signal Processing and Communication Systems*. IEEE. 2010, pp. 1–4.

[43] Akshay Krishna et al. "Sales-forecasting of Retail Stores using Machine Learning Techniques". In: *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*. IEEE. 2018, pp. 160–166.

[44] *Python Data Analysis Library — pandas: Python Data Analysis Library*. URL: `https://pandas.pydata.org/` (visited on 05/21/2019).

[45] *Matplotlib: Python plotting — Matplotlib 3.1.0 documentation*. URL: `https://matplotlib.org/` (visited on 05/21/2019).

[46] *seaborn: statistical data visualization — seaborn 0.9.0 documentation*. URL: `https://seaborn.pydata.org/` (visited on 05/21/2019).

[47] *Project Jupyter*. URL: `https://www.jupyter.org` (visited on 05/21/2019).

[48] *Long Short-Term Memory Networks With Python*. en-US. URL: https://machinelearningmastery.com/lstms-with-python/ (visited on 05/21/2019).

[49] Kedar Potdar, Taher S. Pardawala, and Chinmay D. Pai. "A comparative study of categorical variable encoding techniques for neural network classifiers". In: *International Journal of Computer Applications* 175.4 (2017), pp. 7–9.

[50] Jason Brownlee. *Time Series Forecasting as Supervised Learning*. May 2017. URL: https://machinelearningmastery.com/time-series-forecasting-supervised-learning/.

[51] Agus Widodo, Indra Budi, and Belawati Widjaja. "Automatic lag selection in time series forecasting using multiple kernel learning". In: *International Journal of Machine Learning and Cybernetics* 7.1 (2016), pp. 95–110.

[52] David A. Dickey and Wayne A. Fuller. "Distribution of the estimators for autoregressive time series with a unit root". In: *Journal of the American statistical association* 74.366a (1979), pp. 427–431.

[53] *Time Series Forecasting as Supervised Learning*. URL: https://machinelearningmastery.com/time-series-forecasting-supervised-learning/ (visited on 05/21/2019).

[54] G. Udny Yule. "Why do we sometimes get nonsense-correlations between Time-Series?–a study in sampling and the nature of time-series". In: *Journal of the royal statistical society* 89.1 (1926), pp. 1–63.

[55] M. Kendall and Jean D. Gibbons. "Rank correlation methods edward arnold". In: *A division of Hodder & Stoughton, A Charles Griffin title, London* (1990), pp. 29–50.

[56] Erich L. Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[57] Georgios Drakos. *How to select the Right Evaluation Metric for Machine Learning Models: Part 2 Regression Metrics*. Sept. 2018. URL: https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-2-regression-metrics-d4a1a9ba3d74 (visited on 05/21/2019).

[58] Rob J. Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[59] Christoph Bergmeir and José M Benıtez. "On the use of cross-validation for time series predictor evaluation". In: *Information Sciences* 191 (2012), pp. 192–213.

[60] Barbara Ann Kitchenham, David Budgen, and Pearl Brereton. *Evidence-based software engineering and systematic reviews*. Vol. 4. CRC press, 2015.

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden