

THE DEVELOPMENT OF THE PREDICTIVE MODEL TO FORECAST RETAIL
STORE SALES DURING PEAK AND OFF-PEAK PERIODS

BY

NONGQOTO S

(201636304)

RESEARCH PROJECT

Submitted in partial fulfilment of the requirements for the degree of

HONOURS SCIENCES

(Computer Science)

UNDER THE DEPARTMENT OF COMPUTER SCIENCE

In the

FACULTY OF SCIENCE AND AGRICULTURE

(School of Mathematics and Computer Sciences)

At the

UNIVERSITY OF LIMPOPO,

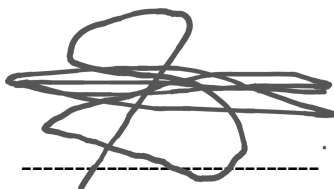
SOUTH AFRICA.

SUPERVISOR: MR J TLOUYAMMA

January 2021

DECLARATION

I declare that the research project titled THE DEVELOPMENT OF THE PREDICTIVE MODEL TO FORECAST RETAIL STORE SALES DURING PEAK AND OFF-PEAK PERIODS under the supervision of Mr J Tlouyamma, is my own work and that all the collaborative contributions and diagrams that were used or quoted in this have been indicated clearly and acknowledged by means of complete references and that this work has not been submitted before for any other degree or professional qualification at any other institution. I am aware of and understand the university's policy on plagiarism.



Signature

Sinethemba Nongqoto

Full names

14 January 2021

Date

ACKNOWLEDGEMENTS

This study would not have been possible without the financial support of the IBM Bursary (International Business Machines Corporation funding). I am especially indebted to Mr Tlouyamma, my project supervisor from the University of Limpopo, in the department of Computer Science under HOD Prof M Velempini, who embraced my career ambitions and who worked actively to provide me with a safe academic time to achieve those goals hence I have a lot of respect for him. He has taught me more than I could ever give him credit for here. He has shown me, by his example, what a good scientist (and person) should be.

I am grateful to all those with whom I have had the privilege of working with during this project.

I also extend my gratitude to my beloved family for the support and encouragements. I would like to thank my mom whom her love, guidance, and prayers are with me in whatever I pursue. And least but no last I would like to pay my greatest gratitude and appreciation to my loving and supportive girlfriend, Deane Chando for her daily encouragement and unending inspiration.

ABSTRACT

Predictive analysis is the use of information, factual equations and machine learning approaches to differentiate the likelihood of future outcomes from data. Sales prediction is an important field in the retail industry, and due to new technologies, a huge amount of attention has increasingly been paid towards improving business operations and profitability. Our study develops the predictive model to forecast retail store sales during peak and off-peak periods. Quantitative research was conducted to evaluate the effectiveness of predictive models in retail settings. Python libraries were used for model development and for data analysis. The sales data was collected from Retail Data Analytics Company with 45 stores located in different regions. Only data from one store was considered for building a model due to it having significant amount of data suitable for building a model. We used sales data collected from 2010 to 2012 to build predictive model. Supervised ML were applied to predict weekly sales. The developed model has shown to have the highest degree of accuracy and has reliably predicted an unseen data or the data it was not trained on.

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

TABLE OF CONTENT

DECLARATION.....	i
ACKNOWLEDGEMENTS.....	ii
ABSTRACT.....	iii
TABLE OF CONTENT	iv
LIST OF FIGURES AND TABLES.....	vi
LIST OF ABBREVIATIONS	vii
CHAPTER 1	1
1.1 INTRODUCTION	1
1.2 PROBLEM STATEMENT.....	2
1.3 MOTIVATION	2
1.4 AIM.....	2
1.5 RESEARCH QUESTIONS.....	3
1.6 OBJECTIVES.....	3
1.7 SCIENTIFIC CONTRIBUTION	3
CHAPTER 2: (LITERATURE REVIEW)	4
CHAPTER 3: (METHODOLOGY)	18
3.1 INTRODUCTION	18
3.2 DATA COLLECTION AND PREPARATION	18
3.2.1 Data Collection.....	18
3.2.2 Modelling and work flows	19
3.3 THE DESCRIPTION OF DATA.....	20
3.4 DATA SPLITTING AND TESTING.....	22
3.5 SOFTWARE TOOLS	22
3.5.1. Python	22
CHAPTER 4: (ANALYSIS OF RESULTS).....	24
4.1 INTRODUCTION	24
4.2 DEALING WITH MISSING VALUES AND DUPLICATE DATA ENTRIES	25
4.3 DATA EXPLORATION	25
4.4 AUTOCORRELATION AND PARTIAL AUTOCORRELATION	26
4.5 FORECAST OF THE STORE-WISE SALES VOLUME	27
4.6 PREDICTIVE POWER FROM EXTERNAL VARIABLES	28
CHAPTER 5 (CONCLUSION AND FUTURE WORK)	34
5.1 CONCLUSION	34

5.2	FINDINGS	34
5.3	FUTURE WORKS	35
6.1	REFERENCES	36

LIST OF FINGURES AND TABLES

Figure / Table	Tittle
Table 1.	Description of data in the store data table
Table 2.	Description of data in the features data table
Table 3.	Description of data in the sales data table
Figure 1.1	System Architecture (Cheriyen, 2018)
Table 4.	Total weekly sales of the top 5 stores
Figure 4.1.	AR Residual R-Squared
Figure 4.2.	Graphical insight of the top 5 stores in terms of sales
Figure 4.3.	Correlation of external variables
Figure 4.4.	AR residuals for weekly sales with/without external variables
Figure 4.4.	The figure shows the run of the actual and predicted data for year 2011
Figure 4.4.	Weekly sales for top 5 stores during holiday/not holiday
Figure 4.8.	Weekly sales for top 5 stores during holiday/not holiday

LIST OF ABBREVIATIONS

ML	Machine Learning
EM	Expected Maximization algorithm
ANN	Artificial Neural Network algorithm
SVM	Support Vector Machine
K-NN	K – Nearest Neighbour
ERP	Enterprise Resource Planning
CRM	Customer relationship Management
S&P	Standard and Band
ARIMA	Autoregressive Blended Moving Average
BPNNs	Back Propagation Neural Networks
MLP	Multilayer Perception
RBFN	Radial Basis Function Network
MAPE	Mean Average Percentage Error
RMSE	Root Mean Squared Error
rANOVA	Repeated measure analysis of variation
APIs	Autoregressive
AR	Application Programming Function
PACF	Partial Autocorrelation Function
MA	Moving Average
SARIMA	Seasonal Autoregressive Integrated Moving Average
VAR	Vector Auto Regression

CHAPTER 1

1.1 INTRODUCTION

Data Science is one of the emerging fields of theoretical and applied research, integrating experience in the field, programming skills and information. Field data science overlaps with business intelligence, a sub-field generally referred to as business analysis. Market analytics closely track historical business records, use statistical mathematics and Machine Learning to model, and forecast the future.

Machine Learning (ML) involves the creation of algorithms and computer programs that help the machine learn its operating environment and predict the future accurately. Machines should be trained using data to learn their environment. The researcher in the field of ML split data into train data set and test data set (Shigeo, 2001). The training data set is mainly used to help the system to understand its operating environment while, on the other hand, the test data set is used to determine the accuracy with which the model can properly fit unseen data. The efficiency of the ML algorithm is dependent on the ability of the computer to predict the outcome at the highest degree of certainty.

“The use of the ML algorithms improves the intelligence of the system” as described by using (Alpaydin, 2004). “ML techniques can be applied to all disciplines and clustering problems. ML algorithms are mainly classified into supervised, unsupervised and semi-supervised” according to (Lytvynenko, 2016). One study noted that “Supervised ML algorithms applies what has been learned in the past to new data using labelled examples to predict future events. Unsupervised ML algorithms are applied when the information used to train is neither classified nor labelled, and semi-supervised ML algorithms consider both supervised and unsupervised learning, since they use both labelled and unlabelled data for training – typically a small amount of labelled data and large amount of unlabelled data” (Expect System team, 2020).

ML techniques were recently, widely used to accurately predict corona virus pandemic that threatened to collapse the economies of the entire world.

Our analysis will apply the ML algorithm to retail store data in order to accurately forecast future sales. The main reason for this research is to assess and analyse the use of ML techniques for the forecasting of sales, to generate models that are detailed and accurate. We can find complicated patterns in sales dynamics by using a supervised machine – learning methods. In this study we will use Machine– Learning Predictive Models, for our analysis we will use stores sales historical data from retail data analytics (Singh, 2017).

1.2 PROBLEM STATEMENT

As one study noted “Retail stores attracts millions of customers yearly and there is a huge competition amongst competitors. Retailers have to adjust business strategies to attract more customers and improve sales. Most of stores opt for markdowns to attract more customers and that may results in cash flow problems in the future. The challenge remains to predict how will the store be affected and to what extent” (Singh, 2017).

1.3 MOTIVATION

Accurately predicting future sales can help any business prevent unexpected cash flow challenges and handle production, personnel and financial needs more efficiently. Sales forecast will make it easier for retail stores to plan their operations more accurately and help them deploy their internal resources with greater efficiency and ultimately acquire the highest investment capital and future growth.

1.4 AIM

The main aim of the study is to develop a model for predicting sales in retail environment.

1.5 RESEARCH QUESTIONS

- What are best methods of predicting sales data?
- Which tools to use to in building predictive models?
- Why is it necessary to build predictive model for retail stores?
- Can machine-learning algorithm be effective in predicting future outcomes?

1.6 OBJECTIVES

This study seeks to achieve the following objectives:

- To evaluate the effectiveness of predictive models in retail settings
- To build accurate predictive model for retail stores
- To predict the department wide sales for each store for the following year
- To provide recommended actions based on the insight drawn, with prioritization placed on largest business impact

1.7 SCIENTIFIC CONTRIBUTION

Predictive analytics not only helps in the development of technically useful models, but also plays an important role in the construction of new hypotheses for further analysis and research. The use of available data to extract inferences and forecasts using predictive analytics has evolved in the industry from a small department in large corporations to an active component in most mid-to large organizations. Sales forecasting can have a huge effect on the success and output of companies. Sales predictions are also a vital aspect of launching a new company with an accurate business plan and better decision-making.

CHAPTER 2: (LITERATURE REVIEW)

Predictive analytics combines a range of statistical techniques from modelling, machine learning and data mining, which analyse current and historical data to determine patterns; predict future outcomes and trends. Predictive analytics does not say what is going to happen in the future, it predicts what could happen in the future with a reasonable degree of reliability. It helps identify potential threats and opportunities for a business. It also helps to better understand consumers, goods, partners and the market. Often there is an uncertain occurrence of significance in the future (Ranjana, 2019).

“Sales forecasting and predicting analytics have a crucial impact on the achievement and performance of companies, companies face numerous challenges concerning accurate forecasts, For instance, they have to place their production plans before exact knowledge about future demands is available” (Thoben, 2015).

“Sales forecasting allow agencies to identify potential problems or risks and design appropriate corrective measures” (Deloitte, 2018). The classification of data is very critical for decision-making. Clustering techniques are very useful in discovering distribution patterns and clustering algorithms employ a distance metric based on similarity measures (Tsai, et al., 2002).

Sales forecasting is the method of using the company's sales records in past years to estimate the company's short-term or long-term sales performance in the future. This is one of the foundations of good financial planning. As in any forecasting method, risk and uncertainty are often inevitable in the forecasting of revenues. Sales forecasting allows companies to produce sufficient goods by predicting consumer demand in advance. Businesses must also make revenue predictions in order to enhance customer loyalty and reduce complaints. They can also boost revenue effectively by handling just-in-time

orders. In order to achieve a high degree of sales forecasting/prediction, it is prudent and strongly recommended that we always look back on what has been achieved in previous studies and study and analyse the steps involved in sales forecasting. Sales forecasting and predicting analytics again have a critical effect on the success and efficiency of businesses. Companies face a variety of problems in terms of reliable predictions. For example, they must put their production plans in place before accurate information of future demands is available (Thoben, 2015). Sales forecasting helps businesses to recognise potential problems or threats and to design effective corrective steps to reduce them (Deloitte, 2018). The classification of data is very critical for decision-making. Clustering methods are very useful in the discovery of distribution patterns and clustering algorithms utilizing distance metric-based similarity measures (Tsai, et al., 2002)

In addition to generating appropriate goods, companies often need to proactively track and maintain inventories. Companies also struggle to track and maintain inventories effectively due to lack of knowledge on potential demand for a commodity. Insufficient inventory control often means that many companies have situations of over-stocking and over-stocking. Sales forecasting helps companies prevent over-stocking and out-of-stock situations by predicting product demand correctly and proactively managing inventory. (bluepiit.com, n.d.)

In general, there are two main types of forecasting approaches to be distinguished: quantitative and qualitative methods. Traditional forecasting is carried out using quantitative, statistical approaches such as time series and regression models, which are the most widely used techniques for predicting sales data according to literature. (Makridakis, et al., 1998).

In different areas of study, it is assumed that the model obtained using neural networks is superior compared to other models. (Lapedes & Farber, 1987)

Published an article in which the network was trained to generate a time series using a specific equation and obtained appropriate results which provided accurate forecasting.

(Thiesing & Vornberger, 1997) Discussed sales forecasting using neural networks trained with a back-spreading algorithm that was used to predict future time series values that consisted of weekly demand for products in the supermarket. Indicators of price promotional campaigns and holidays have been taken into account. The design and implementation of a neural network forecasting framework was identified and built as a prototype for the headquarters of a German supermarket company to help management in the process of determining the expected sales figures. The efficiency of the networks was measured by comparing them to the two prediction techniques used in the supermarket. The comparison showed that neural networks outperformed traditional prediction efficiency techniques. Initially, the exploratory analysis and partial autocorrelation functions were used to evaluate time series sales data in order to verify the presence of seasonal, non-stationary and random components in the data. The number of intermediate layers for the back-propagation model was calculated by the trial and error method. Error analysis of the time series for accurate forecasting was carried out.

(Maita, 2019) Have done Sales prediction using Clustering & ML (ARIMA & Holt's Winter Approach). The author used the sales transaction dataset from UCI ML depository. The dataset contained weekly purchased quantities of 800 products over a year 52 weeks. The cluster partitioning methods were considered helpful in minimizing total intra-cluster variation (Known as total within-cluster variation or total within-cluster sum of square). In preparation to build the predictive model, the author performed product segmentation using clustering. Clusters of items were identified based on their similarities and a common forecast was then computed for each cluster item. There were three popular methods for determining the optimal clusters; those were Elbow

method, Silhouette method and Gap statistic. The study has predicted an average of less than 1.5 transaction per week for items in low demand and maximum average of 24 transactions per week for highly demanded items.

A number of researchers have developed non-linear forecasting models for stock or sales prediction (Chiu, et al., 2008) using improved BPN/Cauchy machine and genetic algorithms to create an effective neural network and forecast Taiwanese electronic stock indexes. Authors, after comparing the back-propagation procedure with other conventional methodologies, found that back-propagation is one of the most appropriate procedures for forecasting, especially non-linear data. Back-propagation improved the weights of each perception to explain the actual results as matching or approaching the expected result.

Another popular form of forecasting is the regression methodology, which focuses on the relationship between sales and all the explanatory factors mentioned above. Regression models investigate how sales can grow if exogenous variables change, e.g. when a marketing action such as a sales promotion campaign is carried out. According to (Mentzer & Moon, 2005) regression models therefore have a large environmental outlook for forecasting sales. Regression models needed large data sets, including past history of and factor and sales volume, and were therefore most useful when the time period is longer than six months (Brannon, 2010).

According to (Sayli, et al., 2016), in order to be competent enough and produce higher sales, business enterprises are continuously searching for a better model or technology for data mining and vital data maintenance. The business sector faces significant challenges in identifying accurate data mining strategies and an efficient predication strategy. Sales data analysis faces many obstacles and core aspects of sales functions include product attribute recognition, price fixation, net sales implementation and the introduction of a new product.

Authors (Sastry, et al., 2013) discussed various prediction methods, sales forecasting strategies and Expectation Maximization (EM) algorithm.

Data mining is the discovery of structures and patterns in large and complex sets (David & Adams, 2015). Appropriate data mining techniques the information from a bulky data set can be converted into a realistic format and can be rendered using supervised and unsupervised learning (Kaur & Mann, 2013). Appropriate revenue prediction methods may be used to make successful business decisions. As suggested by (Korelev & Ruegg, 2015) the error of prediction with the implementation of XGBoost and additional support for the SigOpt Bayesian Optimization Approach can be minimized.

Sales forecasting has been carried out for stores using various data mining techniques. The task involved forecasting sales on any given day in any shop, in order to get acquainted with the task they had previously learned. (Jain, et al., 2015)

According to (Lassen & Vatrapu, 2014) Social media may play a key role in predicting the sales of this model. Although references to social media are the level of exposure that a product receives, Google trends may reflect the interest that potential customers have in the product. Study has shown that search behaviours can reflect purchasing intention and even forecast consumer behaviour and sales of both lower and higher stake purchases (Choi & Varian, 2012).

Previous research has investigated the importance of social media to forecast sales (Asur & Huberman, 2010). The study projected that box office revenues will be amazingly accurate by using several variables such as emotions and the number of tweets in their prediction model. Several other predictive research have adopted the approach taken (Asur & Huberman, 2010). A research based

on this approach by (Lassen & Vatrappu, 2014) projected quarterly iPhone sales by evaluating tweet sentiments and using seasonal tweet weighting to measure the quarterly share of the last calendar year.

(Massaro, et al., 2018) Compared data mining model sales predictive algorithm output based on RapidMiner Workflows. Authors processed the data collection of RapidMiner workflows derived from separate data files and containing information on sales over three years of a wide chain of retail stores. Subsequently, the authors built a deep learning model with a predictive algorithm suitable for sales forecasting. The model was based on an artificial neural network (ANN) algorithm designed to learn the model from historical sales data and pre-processing data. The best-built model used a multilayer neural network along with the "optimized operator" to automatically find the best parameter setting for an implemented algorithm. Other machine learning algorithms have been tested to prove the most effective predictive model. Performance comparisons were made between support vector machine (SVM), k-Nearest Neighbour (k-NN), Gradient Boosted Trees, Decision Trees, and Deep Learning algorithms. The comparison of the degree of correlation between the actual and the expected values, the average absolute error and the relative average error showed that ANN had the best results. The Gradient Boosted Trees approach represented an alternative approach with the second best results. The study was implemented within the context of an industry project aimed at incorporating high-performance data mining models to forecast revenues using enterprise resource planning (ERP) and customer relationship management (CRM) methods.

An improved demand forecasting model using the learning method and a proposed strategy for integration of supply chain decisions by (Kilimci, et al., 2019) has been developed, as demand forecasting is one of the key issues in the supply chains. Research aimed at maximizing stocks, reducing costs, increasing sales, earnings and customer loyalty. Historical data was analysed to enhance demand forecasting by using various approaches, such as machine

learning techniques, time series analysis and interpretation of historical data, using different forecasting methods, including time series analysis techniques, vector regression algorithms and deep learning models. The other innovation of the authors' work was the application of the ensemble strategy to the demand forecasting method by applying a novel decision integration model. The device was implemented and tested on real life data obtained from SOK Market in turkey, which operates as a fast-growing business with 6700 shops, 1500 goods and 23 distribution centres. After a large variety of comparative and detailed trials, the proposed demand forecasting method showed impressive results compared to the state-of-the-art studies. However, unlike the state-of-the-art research, the incorporation of help vector regression, deep learning models and a novel approach for integration into the proposed forecasting method ensures substantial improvement in precision.

(Pavlyshenko, 2018) For their analysis, they used store sales historical data from "Rossman Store Sales" Kaggle competition (Anon., 2018). Calculations were made in the Python environment using the key packages pandas, sklearn, numpy, keras, matplotlib, seaborn.

Jupyter Notebook was used to perform the analysis. In the first place, they carried out a descriptive analysis, which is the study of revenue distributions, of data visualization with different pair plots. Then supervised machine-learning method was considered using historical time series sales and for categorical features, one-hot encoding was implemented when one categorical variable was replaced by n binary variables, while n is the number of unique values of categorical variables. In the forecast, bias on the validation set can be observed which is a constant (stable) under or over-valuation of revenue when the forecast is going to be higher or lower with respect to actual values. The accuracy of the validation set is an important predictor for maximizing the number of iterations of machine learning algorithms. Results have shown that the use of sketching techniques can boost the efficiency of predictive models for the forecasting of sales time series.

According to (Pavlyshenko, 2018) the result of machine-learning generalization is that the regression algorithm captures patterns that occur in all stores or goods. More specific findings can be produced in the case of a short time span. The effect of machine-learning generalization enables predictions to be made in the event of a very limited number of historical sales data, which was important in the new product or store opened. Expert correction can be made by multiplying the forecast by a time-dependent coefficient to take into account transitory processes, e.g. the mechanism of product cannibalization as new products replace other products.

Authors ((Wolpert, 1992); (Rokach, 2018); (Dietterich, 2000); (Rokach, 2005)) -considered the stacking techniques for constructing a range of predictive models. In this approach, the results of the validation set predictions were viewed as input regression for the next stage models. Authors considered a linear model or another form of machine learning algorithm, e.g. Random Forest or Neural Network. On the first stage (Pavlyshenko, 2018) several models were used, most of them based on the XGBoost machine-learning algorithm (Chen & Guestrin , 2016). For the second stacking stage, the author used two models of the Python scikit-learn package: the Extra Tree Model and the Linear Model and the Neural Network Model. The results of the second level were summarized with the third level weights. A number of new features have been constructed, but the most important of them are based on the goal aggregation variable and its lags with the grouping of different variables.

The retention of customers is a big challenge for organizations. . (Mohanty & Ranjana, 2018) Reported that they had no previous customer behaviour study. Organizations were typically faced with difficulties in creating a perfect model and had no perfect method of achieving optimized marketing strategies. In their study, the authors explained how they use predictive analytics using Big Data Analytics and Python software.

There has been substantial research work done by (Rokach, 2005; Rokach, 2005; Simmons, et al., 2010; Dorr & Denton, 2009; Gavrilov, et al., 2000; Kharratzadeh & Coates M, 2012) forecasting the company's stock prices based on an analysis of online media content such as news stories, web blogs, Twitter feeds. For example, (Gavrilov, et al., 2000) Applied data mining techniques on the stock details of different companies by clustering them according to their Standard and Poor (S&P) 500 index, while weblog content was used by (Kharratzadeh & Coates M, 2012) to identify the underlying relationships between the companies to make predictions about the evolution of stock prices. The most notable papers in this regard is from (Asur & Huberman, 2010) it has shown that social media feeds can be used as powerful measures of real-world performance. In their work, they used an analysis of the hourly rate of tweets about movies, re-tweets and sentiment polarity to reliably estimate the revenue of box office movies. In reality, their forecast of film sales based on social media metrics from twitter outperformed the Hollywood Stock Exchange's leading market-based forecasts.

(Penpece & Elma, 2014) Predicted sales using the Artificial Neural Network (ANN) in the food retail sector in Turkey. Since the forecasting of sales quantity and sales revenue is essential for a business to take action for a sustainable completion in the next cycle, it is particularly important for the fast-moving rising industries such as the grocery retail sector. Due to their ability to identify patterns and machine learning, ANN models have been used. The ANN approach was used to estimate sales revenues for the coming year. According to the results, there were high similarities between the forecast and the actual data. The estimated results of their analysis were only 10% larger or smaller than the actual data. Due to this high accuracy, retailers in Turkey may use ANN as a forecasting model.

(Aye, et al., 2013) Forecasted aggregate retail revenues, the case of South Africa aims to boost the ability of portfolio investors to forecast changes in retail chain stock prices. The authors used data from 1970:01 to 2012:05 and from

1987:01 to 2012:05 as an out-of-sample period. They deviated from the uniform symmetrical quadratic loss feature usually used in the forecast assessment of excise duties. Therefore, the authors considered the loss functions of the overweight forecast error in booms and recessions to verify whether a particular model that appears to be a good option on average is indeed preferable in times of economic stress. Authors used the weighted version of the Diebold-Mariano tests to assess the various forecasts. Results have shown that, focusing on a single model alone, their output varies greatly across forecast horizons and across various weighting schemes. In general, however, the combination forecast models provided better forecasts and are generally unaffected by market cycles and time horizons.

Manufacturers need to market their goods. Demand is often seasonal or cyclic. In such situations, understanding how external factors such as costs, weather, the consumer price index and the premium rate could influence the customer's demand for sales will help with the allocation of resources in manufacturing. Predictive analytics take historical sales data and apply forms of regression to forecast future sales based on past sales. In our study we have reviewed that (Maita, 2019) Have done prediction of sales using Clustering & ML (ARIMA & Holt's Winter Approach). The author used the UCI ML repository sales transaction dataset. The dataset included weekly transactions of 800 items over a span of 52 weeks. Cluster partitioning methods were considered to be helpful in reducing total intra-cluster variance (Known as total within-cluster variation or total within-cluster sum of square). In preparation for building a predictive model, the author carried out a segmentation of the product using clustering. Clusters of items were identified on the basis of their similarities and a common forecast was then computed for each cluster item. There were three common methods for determining optimal clusters; Elbow method, Silhouette method, and Gap statistic. The study predicted an average of less than 1.5 transactions per week for products of low demand and a maximum average of 24 transactions per week for items of high demand.

(Omar, et al., 2016) Researchers used the Hybrid Neural Network Model for Sales Forecasting based on ARIMA (Autoregressive Blended Moving Average) and Article Titles search popularity. Authors saw that publishers typically pick appealing titles and headlines for their articles to boost sales, as famous article titles and headlines will encourage readers to buy magazines. Authors' information retrieval methods have been implemented to retrieve terms from the titles of the paper. The popularity indicators of the article titles were then analysed using search indices obtained from the Google search engine. Back Propagation Neural Networks (BPNNs) were successfully used to develop sales forecasting models. The proposed new hybrid neural network model for sales forecasting was focused on the forecast results of time series forecasting and the popularity of article titles, the proposed model used historical sales data, the popularity of article titles, and the result of time series forecasting. Authors used ARIMA forecasting method to learn prediction techniques. The experimental result showed that the proposed forecasting method outperformed conventional techniques which did not consider the popularity of title words.

(Aberg & Dahlen, 2017) Predicted sales to the grocery store department using ML to boost company operations and profitability. The study aimed to compare three ML methods for sales prediction in the food industry: Multilayer Perception (MLP), Support Vector Machine (SVM), and Radial Basis Function Network (RBFN). After these methods were compared due to their prediction accuracy on the daily sales, the performance of the models was determined by using the performance measures: Mean Average Percentage Error (MAPE) and Root Mean Squared Error (RMSE). The end results showed that the SVM performed less error measurements than the other two methods. The difference between the methods was determined by the repeated measure analysis of variation (rANOVA).

The discovery of optimal parameter settings for each model was prioritized due to the time constraints of the analysis. Lack of deeper knowledge about how to

find optimum parameter settings can be seen as a strong limitation in the analysis (Aberg & Dahlen, 2017). Deep knowledge of the features that influence sales will result in a more reliable forecast. The results of a statistical analysis between ML methods could not be very insightful without adequate data.

Predictive research conducted by (Reznek, et al., 2017) on an integrated sales and marketing platform. The strategies implemented in the authors' study allow sales managers and sales representatives to confidently predict sales. The content included a series of slides that included information about a product or service made available by a sales representative to the viewer (e.g. prospective customer) and the content can be shared through a browser-based screen sharing technology that uses scripting machine language codes to identify instances of viewer behaviour. The objective operation of viewer interactions with content was created by computer scripting language codes and automatically uploaded to the analytics platform through one or more application programming interfaces (APIs). The analytical platform applied predictive modelling techniques to objective activity data in order to measure the real contribution of the audience to the content shared by the sales representative. Prediction as to whether potential transactions are likely to close using the historical data set of previous deals that has been made on the analytics platform. Authors also discovered that the automated processing of objective activity data by the analytics platform allows accurate forecasting of sales results by the sales representative and removes subjective human bias from the sales activity data used to predict sales.

“Sales forecasting can be done using different data mining and machine learning techniques where predicting sales on any given day at any store can be carried out”, the detailed analysis and procedures were shown by (Jain, et al., 2015). The tasks involved in forecasting sales on any given day in any store, in order to get acquainted with the task we have previously studied or checked. It seems clear that predictive analysis has had a positive effect on business since the early days. It is also clear that the impact will increase exponentially

as data, models, methods and machine learning continue to evolve on the basis of sophistication, high accuracy and decision-making. The demand for predictive analytics would sweep the entire industry and bring the company to new heights. Market leaders continue to make efforts to transform data and apply analytics with increasing complexity. A brief overview of how to build predictive models in the industry using two of the current algorithms (i.e. time series and logistic regression algorithms) in Python that has been presented. More predictive models can be built based on business scenarios using various tools such as tableau, Python and so on. The performance of these predictive models can be compared to real-time data, i.e. streaming from outside the world to Big Data or batch data. Organizations can use software such as Apache Kafka and Storm for streaming and can use any of the tools in the Hadoop ecosystem framework for batch processing.

For several years, the retail sector has used experience in the past years to come up with new products and schemes. With the aid of predictive analytics, retailers can make decisive action on the basis of real-time data and forecast future trends. After study, they come up with new ideas and offer to attract more customers. It not only helps them recognize the most popular products, but also helps them identify the most popular products or combinations favoured by customers. Also for smaller retailers, integrating these observations with predictive analytics will uncover new future sales and show emerging trends.

By using this model, retailers can estimate the number of items that they will need during peak and off peak period. As a result, the system would allow them to increase their profits.

In our case study, we considered various machine learning approaches for the time series. The approach used by (Pavlyshenko, 2018) and the author's team using a predictive machine-learning model is very similar to the one that we will use to complete our research on retail sales forecasting. We can use retail data mining to reliably forecast future sales. We will be concentrated on discovering

complicated trends in sales dynamics, using supervised machine-learning methods.

The use of regression techniques in sales forecasting can often yield better results than time series methods. ARIMA and ANN models have achieved success in both linear and non-linear domains (Zhang, 2003). One of the key assumptions in regression methodology is that trends in historical data can be replicated in the future. Using staking makes it possible to take into account the variations in outcomes for various models with different sets of parameters and to increase the precision of the validation and out-of-sample data sets.

CHAPTER 3: (METHODOLOGY)

3.1 INTRODUCTION

This section sets out the procedures or techniques used to define, select, process and evaluate information on retail data analytics. Data collection enables an individual or organization to answer specific questions, analyse results, and forecast future probabilities and trends. We use data preparation techniques that will allow us to achieve higher data quality. If the planning and structuring are done, the next step will be the understanding of the data. Once the forecasting models have been created, it will be time to start the training phase. If the model has been tested, it can be used to forecast store sales.

3.2 DATA COLLECTION AND PREPARATION

3.2.1 Data Collection

In this study we consider using data from kaggle (Singh, 2017). The sales related data was collected from 45 stores located in different regions – each store contains a number of departments. The historical dataset to be used in this research is based on Retail Data Analytics Company that maintains historical sales data and the data collection periods ranged from 2010 to 2012. The data are stored in csv files. There are approximately 421 570 sales records contained in these files and occupy about 13MB of storage. The organization holds a variety of promotional trademark activities during the year. These markers precede popular holidays, the four main of which are: Super Bowl, Labour Day, Thanksgiving, and Christmas. Weeks, including such holidays, was weighted five times higher in the evaluation than non-holidays.

3.2.2 Modelling and work flows

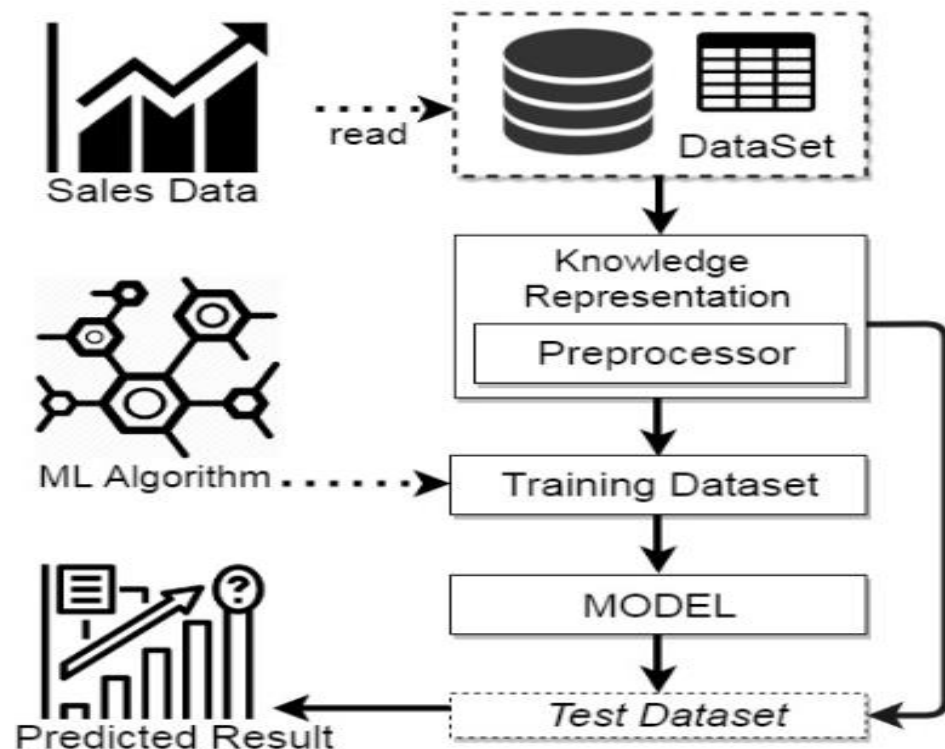


Figure 1.1: System Architecture (Cheriyān, 2018)

Sales Data – Data used to monitor sales. Sales planning data such as industry data used to produce sales forecasts.

Data Set – The data set includes sales information for the development of our model.

It is a set of data arranged as a table, rows and columns.

Knowledge Presentation – It is a fundamental stage for data analysis and knowledge discovery, and hence we consider the pre-processing stage to be critical for knowledge discovery and to have a major impact on predictive accuracy. We can transform pre-processed ML ready data using scaling, attribute decomposition, and attribute aggregation.

Training Dataset – In ML, the common task is to research and build algorithms that can learn from and predict data. The model is trained in a training dataset using a supervised learning process. The training dataset is mainly used to help the system to understand its operating environment.

Model – After specifying the variables that we will use for the analysis, it will be time to use a neural designer to create a predictive model for store sales.

Test Dataset – The dataset used to provide an unbiased assessment of the final model fit to the training dataset. Before using the model to forecast sales, the last step would be to assess its predictive capacity on an isolated collection of data that has not been used before for testing. It is used to test the accuracy with which the model will match the unseen data properly.

Predicted Result – The final effects of the revenue data for the future are based on historical data.

3.3 THE DESCRIPTION OF DATA

In this section, we describe data contained in three data tables:

- Table 1: Anonymised information on the 45 stores, showing the type and size of the store. The table includes 45 rows and 3 columns.
- Table 2: Additional data belonging to the store, department, and regional activities for the given dates. The table consists of 8190 rows and 12 columns.
- Table 3: Historical sales data, which covers from 2010 to 2012 within this tab. The table consists of 421570 rows and 5 columns.

Table 1: description of data in the store data table

Variable	Description	Data Type
Store	1 – 45 Stores (Number of the store)	Integer
Type	A,B, and C	Object
Size	Size of the store	Integer

Table 2: description of data in the features data table

Variable	Description	Data Type
Store	the store number	Integer
Date	The week	Date
Temperature	Average temperature in the region	Float
Fuel Price	Cost of the fuel in the region	Float
Markdown1-5	Anonymised data related to promotional markdowns. Markdown data is only available after November 2011, and is not available for all stores all the time. Any missing value is marked with a NaN	Float
CPI	The consumer price index	Float
Unemployment	The unemployment rate	Float
IsHoliday	Whether the week is a special holiday week	Boolean

Table 3: description of data in the sales data table

Variable	Description	Data Type
Store	The store number	Integer
Dept	The department number	Integer
Date	The week	Date
Weekly_Sales	Sales for the given department in the given store	Float
IsHoliday	Whether the week is a special holiday week	Boolean

In our analysis, because the data given is in multiple files, some of the columns are present in more than one file. We join and merge DataFrames since Joining and merging DataFrames is a key process to start with data analysis and machine learning tasks. It is one of the toolkits that any Data Analyst or Data Scientist should learn because, in almost all cases, data comes from multiple sources and files. We need to put all the data to one place by merging and

some kind of logic to join, and then we'll start with our analysis. In our research, we will concentrate more on the top 5 most profitable stores to better understand our sales data and sales results across peak and off peak times.

3.4 DATA SPLITTING AND TESTING

The use of ML algorithms increases machine intelligence without human interference. "ML is used to optimize the performance criterion by using sample data or past experience.

Next, we use the descriptive analytics, which is the sales distribution analysis, the data visualization with different pair plots. It would be useful to find the similarities and sales factors to concentrate on. In the case of a small pattern, we can find bias by using a linear regression on the validation collection. We will consider the supervised approach to Machine-Learning using historical time series sales. We will use the Random Forest algorithm for the case study (Breiman, 2001). The accuracy of the testing set is an important predictor for maximizing the number of iterations of machine learning algorithms. The impact of machine-learning generalization would allow us to make predictions in the event of a very limited number of historical sales data, which is crucial when a new product or store is launched. One of the key assumptions of the regression approach is that trends in historical data will be replicated in the future.

3.5 SOFTWARE TOOLS

3.51. Python

Python is used for developing predictive model and data analyses. The benefits of using the python environment are its simplicity, low learning curve, well supported and recorded. It has been commonly used in academic and industrial circles. Python has a range of useful analytical libraries that will be applied to the predictive model creation process. Python can also be used for data

cleaning to handle common issues such as conflicting column names, missing data, and various data types and duplicate rows, etc.

We installed anaconda which is the distributor for python and then run python 3 in the JUPYTER notebook using Windows 10 64-bit operating system, Installed RAM 8 GB and Intel(R) Core i5-6300U. We also imported necessary python libraries for data manipulation and data analysis. Database querying tool import data manipulation tools such as pandas and numpy. The imported libraries help to create data statistics and clearly visualised data to gain better insight and clearly explain the trends. Seaborn and matplotlib.pyplot libraries are used for data visualization.

CHAPTER 4: (ANALYSIS OF RESULTS)

4.1 INTRODUCTION

Predictive analytics in data science relies on the explanatory data analysis, which is exactly what we discussed in our previous chapters. This section sets out the procedures or techniques used to define, choose, process and evaluate information on retail data analytics. We used data preparation techniques to achieve higher data quality. Once the planning and structuring were done, the next step was to understand the data. Data is important for a high-performance marketing team.

In this study we considered data from kaggle (Singh, 2017). The sales related data was collected from 45 stores located in different regions – each store contains a number of departments. The historical dataset used for this study is based on retail data analytic company that maintains historical sales data. The data collection periods ranged from 2010 to 2012. The data was stored in csv files and files contained approximately 421 570 sales records occupy about 13MB of computer's memory. The company runs several promotional markdowns events throughout the year. These markdowns precede prominent holidays, the four largest of which are: Super Bowl, Labour Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holidays.

A brief descriptions of tables containing data are as follow:

- Store Table: Anonymised information about the 45 stores, indicating the type and the size of the store. The table contain 45 rows and 3 columns.
- Features Table: Additional data related to the store, department, and regional activity for the given dates. The table contains 8190 rows and 12 columns.
- Sales Table 3: Historical sales data, which covers from 2010 to 2012 within this tab. The table contains 421 570 rows and 5 columns.

We used python for our analysis of results since it is most popular tool across all data science sub-disciplines. The advantage of using python is that it can integrate

and manipulate data with ease. It also has built-in libraries for building models and support for mathematical and statistical computations.

4.2 DEALING WITH MISSING VALUES AND DUPLICATE DATA ENTRIES

We merged our data frame sales data with features data on “Store”, “Date”, “IsHoliday” using left join and then stores data on “Store”, for missing values we put 0 to all NaN and converting temperature to degrees celsius. We checked data and features training data to drop duplicates values. After data cleaning, we gained some insight into data with external variables by plotting some of the variables present in the data frame such as temperature, fuel price, CPI (the consumer price index), unemployment rate.

4.3 DATA EXPLORATION

In Table 4, we listed top 5 performing stores and explored sales data. The stores with highest weekly sales were considered for building predict model for the future sales.

Store	Weekly_Sales
20	3.013978e+08
4	2.995440e+08
14	2.889999e+08
13	2.865177e+08
2	2.753824e+08

Table 4. Total weekly sales of the top 5 stores

We used descriptive analysis to analyse the sales distribution. It is crucial to find correlations and sales drivers on which to focus on. In case of small trend, we can find bias using linear regression on the validation set. Supervised ML approach

was considered using sales historical time series. The effect of ML generalization enabled us to make prediction in case of the very small historical sales dataset, which is important when a new product or store is launched. One of the main assumptions of regression methods is that the patterns in the historical data will be repeated in future.

4.4 AUTOCORRELATION AND PARTIAL AUTOCORRELATION

Identification of an AR (Autoregressive) model is often done with PACF (Partial autocorrelation function).

- For an AR model, the theoretical PACF "shuts off" past the order of the model. The phrase "shuts off" means that in theory the partial correlations are equal to 0 beyond that point. Put another way, the number non-zero partial autocorrelations gives the order of the AR model. By the "order of the model" we mean the most extreme lag of x that is used as a predictor.

Identification of an MA (Moving Average) model is often done with the ACF (Autocorrelation function) rather than the PACF.

- For an MA model, the theoretical PACF does not shut off, but instead tapers toward 0 in some manner. A clearer pattern for an MA model is the ACF. The ACF will have non-zero autocorrelations only at lags involved in the model.

Given the seasonality observed from the ACF and the PACF function, the AR model is implemented including seasonality from weeks.

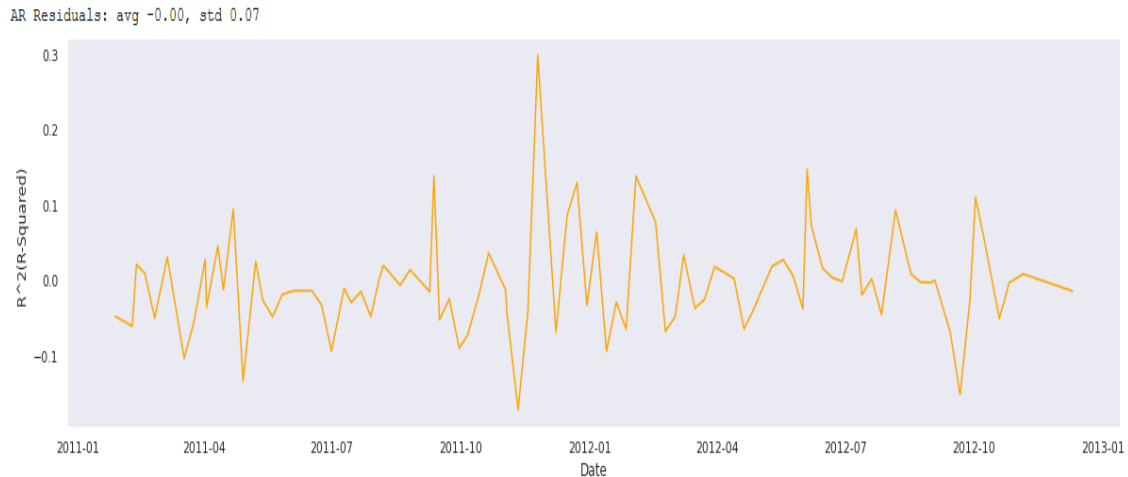


Figure 4.1 AR Residual R-Squared

In figure 4.1 the predictive model has an R^2 score factor of 0.41 (max score for perfect prediction would be 1). The residual distribution was centred in 0 with an STD of 7%. The r-squared reveals that 41% of the data fit the regression model.

4.5 FORECAST OF THE STORE-WISE SALES VOLUME

We used the following key packages for graphing and data analysis: pandas, sklearn, numpy, keras, matplotlib, seaborn. These libraries played a crucial role in the design of our predictive model and analysis of the data in our study. Figure 4.2 shows plotted data from selected top five stores upon which the model will be developed. For simplicity, only store 20 data was used to develop a predictive model which could be in future be rolled out to the four remaining stores. The choice was purely made based on the sufficient data store 20 had.



Figure 4.2: Graphical insight of the top 5 stores in terms of sales

The figure compares sales data for five stores from 2010 to 2012. It is a common trend that all stores experience high sales around December holidays. Sales increase during the peak period and drops considerably during off-peak periods. It can be observed from the figure that January had steep decline in sales. Sales started picking again towards Easter holidays. Store 14 had better sales until 2011 and performed poorly throughout 2012. Store 4 and store 20 performed exceptionally well from early 2011 to the end of 2012. Store 2 was the least performer compared to other stores as it can be observed from the figure.

4.6 PREDICTIVE POWER FROM EXTERNAL VARIABLES

From figure 4.3 the external variables had some correlation with the 1-day lagged sales time series. This means that they had some degree of predictive power at 1 day and can be used to improve our model. The 'MarkDown' and the 'Temperature' being the most correlated and anti-correlated variables respectively. It shows that markdown and temperature have the most impact on weekly sales as the external variables. Sales performance decreases due to

bad temperatures, since most consumers are in doors, while trademarks help to boost sales.

```
shifted_sales    1.000000
MarkDown5        0.084797
MarkDown2        0.050189
MarkDown1        0.035654
CPI              0.021002
MarkDown4        0.013042
MarkDown3        0.002624
Fuel_Price       -0.033798
Unemployment     -0.067482
Temperature      -0.154182
Name: shifted_sales, dtype: float64
```

The external variables available have some correlation with the 1-day lagged sales time series. This means that they have some degree of predictive power at 1 day and can be used to improve our model. The 'MarkDown' and the 'Temperature' being the most correlated and anti-correlate variables respectively.

Figure 4.3 Correlation of external variables

Figure 4.4 presents the results of the model and shows that the external variables had a potential to improve the accuracy of the prediction by more than 40% (R^2 score: 0.58% w.r.t 0.34). The standard deviation of the residual improve by about 30% (7% w.r.t. 8%).

Since time series forecasting is one of the major building blocks of ML. There were many methods in the literature to achieve the same results; for example, Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving-Average (SARIMA), Vector Auto regression (VAR), and so on. In our study we used the regression approaches for sales forecasting since they often give better results than time series methods.

AR Residuals: avg -0.01, std 0.08
 AR with Ext Residuals: avg -0.00, std 0.07

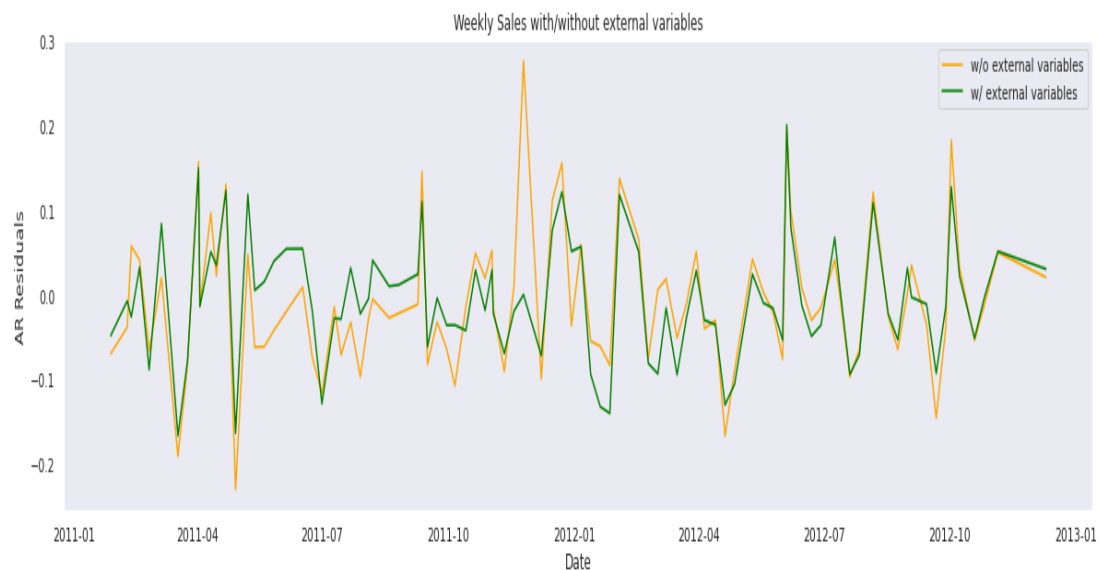


Figure 4.4: AR residuals for weekly sales with/without external variables

ARIMA and ANNs (Artificial neural networks) models have achieved successes in both linear and nonlinear domains (Zhang, 2003). One of the main assumptions of regression methods is that the patterns in historical data will be repeated in the future. Using staking makes it possible to take into account the differences in the results for multiple models with different sets of parameters and improve accuracy on the validation and on the out-of-sample data sets.

In the figure 4.5, we predicted weekly sales on the store 20 since is the most performing store among other stores. We have managed to check how the predicted data would have been from the week 54 (which is the 2nd week of the year 2011) to week 104. Weekly sales from the figure shows an actual data and forecast represent the predicted data. It can be observed that actual and predicted data have the same patterns. Hence our model has the higher degree of accuracy and it can reliably predict unseen dataset.

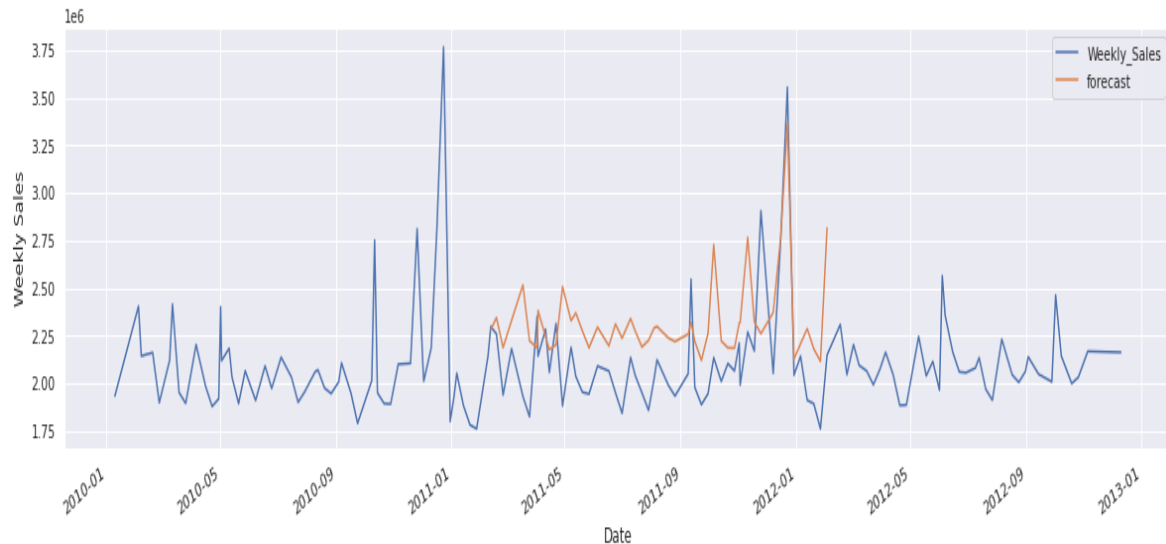


Figure 4.5: The figure shows the run of the actual and predicted data for year 2011

Predictive analytics take historical sales data and applies forms of regression to predict future sales based upon past sales. Good predictive models find additional factors that influenced sales in the past and apply those factors to forecasted sales models.

Back Propagation Neural Networks (BPNs) was successfully used to develop models for sales forecasting. Author (Maita, 2019) have used ARIMA forecasting method to learn prediction techniques.

Figure 4.6 shows historic sales data pattern and the pattern in predicted or forecasted data. The predictive model was developed from three years of sales data (2010 to 2012) to predict sales in 2013. The patterns in the real and predicted sales data have the same trends. This shows that the model is accurate and correctly fit the real data it was trained on. We applied techniques used in the supervised ML to predict weekly sales.

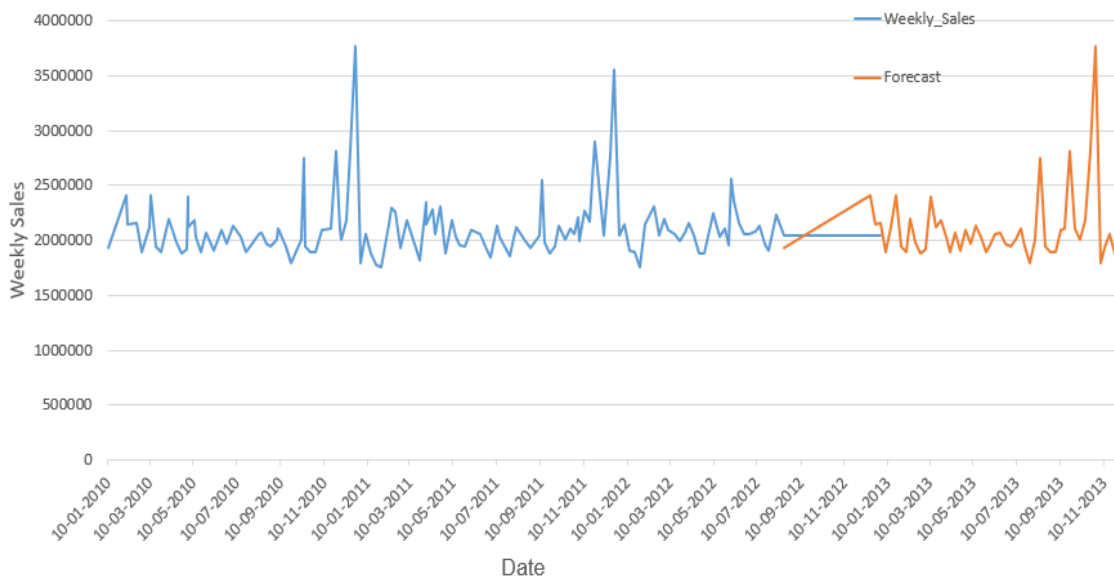


Figure 4.6: Future weekly sales for a year ahead (2013)

The forecasted weekly sales look similar to the historical weekly sales data. The weekly sales are expected to increase around the end of the year and December holidays, so the retail store company should be well prepared and make sure there is enough stock and production.

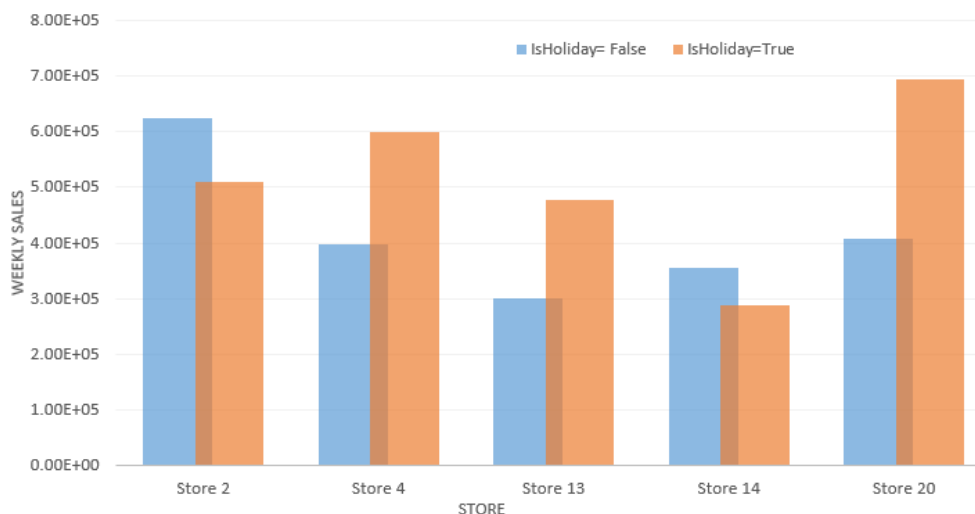


Figure 4.7: Weekly sales for top 5 stores during holiday/not holiday

Sales forecasting is a strategy that helps companies to predict and measure the income. It uses past demand to determine how much inventory departments need to stock up to meet the demands. Figure 4.7 compares sales during peak and off-peak period. And by looking at these, we can confidently say that Markdown that precede holidays has a great impact on the retail stores as they

increase sales. We can see that store 20 had the highest sales during holidays among 4 of other stores, while store 2 had the highest sales when it was not a holiday. The holiday sale always attracts customers.

CHAPTER 5 (CONCLUSION AND FUTURE WORK)

5.1 CONCLUSION

Sales forecasting is very important for any organization, particularly for the big ones. This method is very challenging because there are a lot of variables that need to be taken into account. In order to meet the attainable targets and to achieve them effectively, businesses are keen to foresee future sales times. In our study it is a common trend that all stores experience high sales around December holidays. Sales increase during the peak period and drops considerable during off-peak periods. It was observed that January had steep decline in sales. Sales started picking again towards Easter holidays. We also seen that stores have more sales during a holiday compared when there is no holiday and also external variables have an impact on weekly sales and they improve our predictions by more than 40%. The projected findings are consistent with the presumption or regression that says "the trends in historical data will be replicated in the future." the supervised ML was applied to predict weekly sales hence the predictive model showed a high degree of accuracy as it perfectly fit the historical (real) data it was trained on.

5.2 FINDINGS

From the predicted weekly sales for retail data analytics, it was observed that the predictive model forecasted retail revenues with high degree of accuracy. The weekly sales are expected to increase very high around the end of the year and December holidays. Sales forecasting plays a pivotal role in the financial planning of business for any organization. Financial and sales planning with the help of the sales forecasts improves the performance and decision making of businesses as it greatly help get the information needed as well as the profit. Python libraries were used for model development and for data analysis.

5.3 FUTURE WORKS

The retail sector is changing through a time of unparalleled transition. Emerging technologies continue to cause digital disruption, and customer service is quickly becoming a new currency. Retailers also incorporate data analytics into every stage of their operation, including revenue forecasts, optimization of stores and product recommendations. As customers are accustomed to the effect of these new technologies on their shopping experience, their aspirations are rising higher than ever before.

- The ability of retailers to successfully use AI, data analytics and other new technologies to meet changing consumer demands will be a key determinant of success in the next decade. These innovations would also have a dramatic impact on organizational operations, such as staff management, inventory and sustainability initiatives.

6.1 REFERENCES

- Aberg, R. & Dahlen, C., 2017. Predicting sales in a food store department using machine learning. *Forsaljningsprediktion i en matvarubutik med hjalp av maskinlarning*, 12 June.
- Alpaydin, E., 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning). *The MIT Press*.
- Anon., 2018. *Kaggle.com*. [Online]
Available at: <http://www.kaggle.com/c/rossmann-store-sales>
- Asur, S. & Huberman, B. A., 2010. "Predicting the future with social media" in *Web Intelligence and Intelligent Agent Technology (WI-IAT)*. s.l., IEEE, pp. 492-499.
- Aye, G. C., Balcilar, M., Gupta, R. & Majumdar, A., 2013. Forecasting Aggregate Retail Sales: The Case of South Africa. *University of Pretoria; Department of Economics Working Paper series*, February.
- Batista, M., n.d. *Estimation of the final size of the second phase of coronavirus COVID-19 epidemic by the logistic model*. s.l., s.n.
- bluepiit.com, n.d. Importance Of Sales Forecasting For Businesses. <https://www.bluepiit.com/blog/importance-sales-forecasting-businesses/#>.
- Brannon, E., 2010. *Fashion forecasting (3rd ed)*. New York: Fairchild Books.
- Breiman, L., 2001. Random forests. *Mach. Learn*, pp. 45, 5-32.
- Bronshetein, A., May 17, 2019. *Train/Test Split and Cross Validation in Python*. s.l., towards data science.
- Chen, T. & Guestrin, C., 2016. *Xgboost: A scalable tree boosting system*. San Francisco, CA, USA; ACM: New York, NY, USA, s.n., pp. 785-794.
- Cheriyian, S., 2018. *Intelligent-Sales-Prediction-Using-Machine-Learning-Techniques*. s.l., www.researchgate.net.
- Chiu, D. Y., Chen, T. S. & Pan, Y. C., 2008. Using improved BPN/Cauchy machine and genetic algorithms to build an efficient neural network and to forecast Taiwanese electronic stock indexes. *Journal of Financial Studies*, 16(no. 4), p. 213.

Choi, H. & Varian, H., 2012. Predicting the present with Google trends. *Economic Record (88: special issue SI)*, pp. 2-9.

David, J. H. & Adams, M. N., 2015. Data Mining.

Deloitte, 2018. *Sales Forecasting*. s.l.:Alfredo Maria Garibaldi, Daniele Pier Giorgio Bobba, Marco Leani, Alberto Ferrario.

Dietterich, T. G., 2000. *Ensemble methods in machine learning*. Italy, Switzerland, Cagliari, Cham, pp. 1-15.

Dorr, D. H. & Denton, A. M., 2009. Establishing relationships among patterns in stock market data. In: *Data & Knowledge Engineering*. s.l.:s.n., pp. 318-337.

Expect System team, 2020. *Expect System*. [Online] Available at: <http://expertsystem.com/machine-learning-definition/> [Accessed 17 September 2020].

Fong , S. J., Li, G., Dey , N. & Herrera-Viedma, C. R., 2019. *Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak*. s.l., s.n.

Friedman, J. H., 2001. Greed function Approximation: A gradient boosting machine. *Ann. Stat*, pp. 29, 1189-1232.

Friedman, L. H., 2002. Stochastic gradient boosting. *Compu. Stat. Data Anal*, pp. 38, 367-378.

Gavrilov, M., Anguelov, D., Indyk, P. & Motwani, R., 2000. Mining the stock market (extended abstract): which measure is best? . In: *Mining the stock market*. s.l.:ACM,2000,edn, pp. 487-496.

Jain, A., Menon, M. N. & Chandra, S., 2015. *Sales Forecasting for Retail Chains*. s.l., s.n.

Jain, A., Menon, M. N. & Chandra, S., n.d. Sales Forecasting for Retail Chains.

James, G., Witten, D., Hastie, T. & Tishirani, R., 2013. An Introduction to Statistical Learning. *Cham*, Volume 112.

Kaur, N. & Mann, A. K., 2013. Review paper on clustering techniques. *Global Journal of Computer Science and Technology*.

Kharrazzadeh, M. & Coates M, 2012. Weblog Analysis for Predicting Correlations in Stock Price Evolutions. In: *Weblog Analysis for Predicting Correlations in Stock Price Evolutions*. s.l.:edn.

Kilimci, Z. H. et al., 2019. An improved demand forecasting model using learning approach and proposed decision integration strategy for supply chain. *Research Article*, 26 March, p. 15.

Korelev, M. & Ruegg, K., 2015. Gradient Boosted Trees to Predict Store Sales.

Lapedes, A. & Farber, R., 1987. How neural nets work. In: *Proceedings of the 1987 International Conference on Neural Information Processing Systems*. s.l.:s.n., pp. 442-456.

Lassen, N. B. & Vatrupu, R., 2014. *Predicting iphone sales from iphone tweets*. s.l., doi:10.1109/edoc.2014.20.

Lytvynenko, T. I., 2016. Problem of data analysis and forecasting using decision trees method.

Maita, S., 2019. *towards data science (Sales Prediction using Clustering & Machine Learning (ARIMA & Holt's Winter Approach)*. [Online] Available at: <http://towardsdatascience.com/clustering-machine-learning-combination-in-sales-prediction-330a7a205102> [Accessed 18 October 2019].

Makridakis, S., Wheelwright, S. C. & Hyndman, R. J., 1998. *Forecasting: Methods and Applications*, s.l.: Wiley.

Massaro, A., Maritati, V. & Galiano, A., 2018. DATA MINING MODEL PERFORMANCE OF SALES PREDICTIVE ALGORITHMS BASED ON RAPIDMINER WORKFLOWS. *Dyrecta Lab, IT research Laboratory, via Vescovo Simplicio, 45 70014 Conversano (BA), Italy*, Volume 10 .

Mentzer, J. T. & Moon, M. A., 2005. Sales Forecasting Management: A Demand Management Approach. In: *SAGE Books*. s.l.:SAGE Publications, Inc.

Mohanty, A. & Ranjana, P., 2018. *A Framework For Effective Processing Of Jobs In Hadoop*. s.l., s.n.

- Omar, H., Hoang, V. H. & Liu, D.-R., 2016. Research Article. In: S. Sanei, ed. *A Hybrid Neural Network Model for Sales Forecasting Based on ARIMA and Search Popularity of Article Titles*. Taiwan: Hindawi Publishing Corporation, p. 9.
- Pavlyshenko, B., 2018. *Machine-Learning Models for Sales Time Series Forecasting*. Lviv, Ukraine.
- Pavlyshenko, B., 2018. *Machine-Learning Models for Sales Time Series Forecasting*. Lviv, Ukraine.
- Penpece , D. & Elma, O. E. E., 2014. Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry. *International Journal of Trade, Economics and Finance*, 5(5).
- Ranjana, A. M., 2019. *Usage of Predictive Research on further Business*. s.l.:International Journal of Innovative and Exploring Engineering (IJITEE).
- Reznek, D. A. et al., 2017. *PREDICTIVE ANALYTICS IN AN AUTOMATED SALES AND MARKETING PLATFORM*. United States, Patent Application Publication.
- Rokach, L., 2005. Ensemble methods for classifiers. In: *Data Mining and Knowledge Discovery Handbook*. Cham, Switzerland: Springer, pp. 957-980.
- Rokach, L., 2018. Ensemble-based classifiers. *Artificial Intelligence*, pp. 1-39.
- Sastry, S. H., Babu, P. & Prasada, M. S., 2013. Analysis & Prediction of Sales Data in SAP-ERP Systems using Clustering Algorithms.
- Sayli, A., Ozturk, I. & Ustunel, M., 2016. Brand loyalty analysis system using K-Means algorithm. *Engineering Technology and Applied Sciences*.
- Shigeo, A., 2001. Generation of Training and Test Data Sets. In: *Pattern Classification*. s.l.:Kobe University, pp. 239-247.
- Simmons, B. E., Huffaker, M. P., Teng, C. & Adamic, L., 2010. *The social dynamics of economic activity in a virtual world*. [Online] Available at: <http://misc.si.umich.edu/publications/18>
- Singh, M., 2017. *Kaggle: Retail Data Analytics*. [Online] Available at: <https://www.kaggle.com/manjeetsingh/retaildataset> [Accessed 22 July 2020].

Thiesing, F. M. & Vornberger, O., 1997. Sales Forecasting Using Neural Networks. 9-12 June, 4(IEEE), pp. 2125 - 2128.

Thoben, K.-D., 2015. A survey on retail sales forecasting and prediction in fashion markets. In: *Systems Science & Control Engineering*. Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK: Taylor & Francis, pp. 154-155.

Thoben, K.-D., 2015. A survey on retail sales forecasting and prediction in fashion markets. In: *Systems Science & Control Engineering*. Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK: Taylor & Francis, pp. 154-155.

Tsai, C. F., Wu, H. C. & Tsai, C. W., 2002. "A new data clustering approach for data mining in large databases", *Parallel Architectures Algorithms and Networks 2002. I-SPAN'02. International Symposium*, pp. 315-320.

Tsai, C. F., Wu, H. C. & Tsai, C. W., 2002. "A new data clustering approach for data mining in large databases", *Parallel Architectures Algorithms and Networks 2002. I-SPAN'02. International Symposium*, pp. 315-320.

Wolpert, D. H., 1992. Stacked generalization. *Neural Network*, May, pp. 241-259.

Zhang, P. G., 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* , Volume 50, pp. 159-175.