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

BY

STUDENT NAME

NONGQOTO S

STUDENT NUMBER

201636304

A RESEARCH PROPOSAL SUBMITTED FOR THE HONOURS OF COMPUTER SCIENCES IN THE DEPARTMENT OF COMPUTER SCIENCES

SCHOOL OF MATHEMATICAL AND COMPUTER SCIENCES, FACULTY OF SCIENCE AND AGRICULTURE, UNIVERSITY OF LIMPOPO, SOUTH AFRICA.

SUPERVISOR: MR J TLOUYAMMA

YEAR: 2020

Table of Contents

[**1.** **Introduction** 3](#_Toc51684188)

[**2.** **Problem Statement** 4](#_Toc51684189)

[**3.** **Motivation** 4](#_Toc51684190)

[**4.** **Aim** 4](#_Toc51684191)

[**5.** **Research Questions** 4](#_Toc51684192)

[**6.** **Objectives** 4](#_Toc51684193)

[**7.** **Literature Review** 5](#_Toc51684194)

[**8.** **Methodology** 7](#_Toc51684195)

[**8.1** **Introduction** 7](#_Toc51684196)

[**8.3. The description of data** 8](#_Toc51684197)

[**8.4** **Software tools** 11](#_Toc51684198)

[**9.** **Scientific contribution** 11](#_Toc51684199)

[**10.** **References** 11](#_Toc51684200)

[I. References 11](#_Toc51684201)

# **Introduction**

Data Science is one of the emerging areas of theoretical and applied research and it combines domain expertise, programming skills and knowledge. The field data science overlaps with business intelligence, which has a sub field commonly referred to as business analytics. Business analytics closely observe the historical data of the business, use statistical mathematics and Machine Learning to modelling, and predict the future.

Machine Learning (ML) involves the development of the algorithms and computer programs that help the machine to learn its operational environment and accurately predict the future. In order for machines to learn their environment, they should be trained using data. 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 allow the machine to understand its operating environment while on the other hand, test data set is used evaluate the accuracy at which the model can properly fit unseen data. The effectiveness of the ML algorithm is based on the machine’s ability to predict the outcome at the highest level 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 manly classified into supervised, unsupervised and semi-supervised (Lytvynenko, 2016). 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 study will apply ML algorithm to retail store data to accurately predict future sales. The foremost reason of this research is to evaluate and analyse the use of ML techniques for sales forecasting, to produce models which are comprehensive and reliable. We can find complicated patterns in the sales dynamics, using supervised machine – learning methods. Some of the most popular are tree-based ML algorithms (James, et al., 2013), e.g., Random Forest (Breiman, 2001), Gradient Boosting Machine (Friedman, 2001) and (Friedman, 2002). 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).

# **Problem Statement**

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).

# **Motivation**

Accurately predicting future sales and production can help any company to avoid unforeseen cash flow problems and manage production, staff and financial needs more effectively. Sales prediction can allow a retail stores to manage business more effectively and help them allocate its internal resources with greater efficiency and ultimately acquire the highest investment capital and future growth.

# **Aim**

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

# **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?

# **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
* To investigate the effects of ML in predicting the future outcomes given retail data.

# **Literature Review**

Sales forecasting and predicting analytics have a crucial impart on the achievement and performance of companies. Companies faces 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). Classification of data is very essential in 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).

(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.

There has been previous research that explored the relevance of social media to predict sales, such as the work of (Asur & Huberman, 2010) who predicted box office sales remarkably accurate by including many variables such as sentiments and the frequency of tweets into their prediction model. Various other prediction research has followed Asur and Huberman’s (2010) approach. A study based on this method by Lassen et al. (2014) predicted quarterly iPhone sales by analysing the sentiments of tweets and using a seasonal weighting of tweets to calculate the given quarter’s proportion of the last calendar year.

And the has been another substantial research work done by (Simmons, et al., 2010) (Dorr & Denton, 2009) (Gavrilov, et al., 2000) (Kharratzadeh & Coates M, 2012) in the direction of predicting the stock prices of the companies based on the analysis of content from the online media such as news items, web blogs, twitter feeds. For example, (Gavrilov, et al., 2000) applied data mining techniques on the stock information from various companies by clustering them according to their Standard and Poor (S&P) 500 index, whereas the content from the weblogs 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) showed that social media feeds can be used as effective indicators of the real-world performance. In their work, they used analysis of hourly rate of tweets about movies, the re-tweets and sentiment polarity to accurately forecast the box-office movies revenue. In fact, their prediction of movie revenues based on the social media measures from twitter outperformed the leading market-based predictions of the Hollywood Stock Exchange.

(Pavlyshenko, 2018)For their analysis, they used store sales historical data from “Rossman Store Sales” Kaggle competition (Anon., 2018). The calculations were conducted in Python environment using the main packages pandas, sklearn, numpy, keras, matplotlib, seaborn. To conduct the analysis, Jupyter Notebook was used. They firstly conducted the descriptive analytics, which is the study of sales distributions, data visualization with different pair plots. Then supervised machine-learning approach it was considered using sales historical time series and for categorical features, one-hot encoding was applied when one categorical variable was replaced by n binary variables, whereas n is the amount of unique values of categorical variables. In the forecast, bias on validation set may be observed which is a constant (stable) under or over-valuation of sales when the forecast is going to be higher or lower with respect to real values. The accuracy on the validation set is an important indicator for choosing an optimal number of iterations of machine-learning algorithms. The results showed that using staking techniques, can improve the performance of predictive models for sales time series forecasting.

(Aberg & Dahlen, 2017) Predicted sales in a food store department using ML in order to improve business 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 lower error measures than the other two methods. The difference between the methods was determined by the repeated measure analysis of variation (rANOVA).

Finding of optimal parameter settings for each model were down prioritized due to the time constraints of the study. Lack of deeper knowledge in how to find optimal parameter settings can be seen as a strong limitations of the study of (Aberg & Dahlen, 2017). Deep understanding of what features that affects the sales would result in a more accurate prediction. The results of a statistical comparison between ML methods might not be very informative without sufficient amount of data.

# **Methodology**

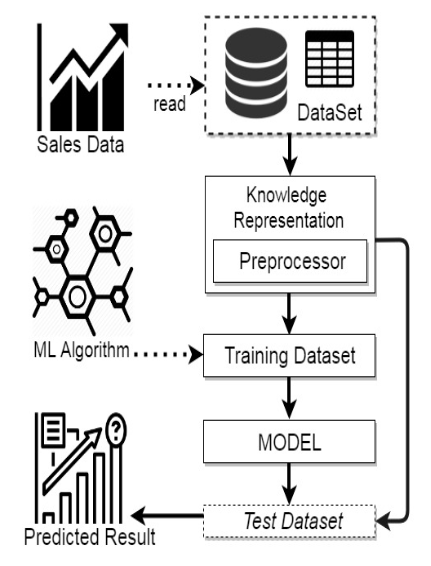
# **Introduction**

In this section we will specify the procedures or techniques used to identify, select, process and analyse information about retail data analytics. Data collection enables a person or organization to answer relevant questions, evaluate outcomes and make predictions about future probabilities and trends. We use data preparation techniques that will allow us to achieve higher data quality. Once preparation and structuring are completed then the next will be data understanding. Once the forecasting models are developed, it will be time to start the training process. Once the model has been tested, it can be used to predict the sales of the stores.

* 1. **Data Collection and Preparation**
     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 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.

* + 1. Modelling and work flow



*Figure 1: System Architecture* (Cheriyan, 2018)

**Sales Data** – Information that is used to manage sales. Sales planning data such as market data that is used to generate sales forecasts.

**Data Set** – The data set contains sales information for creating our model. It is the collection of data structured as a table, in rows and columns.

**Knowledge Presentation** – It is fundamental stage for data analysis and knowledge discovery, therefore we consider the pre-processing stage as an important for knowledge discovery and has a significant impact on predictive accuracy. We will transform pre-processed data ready for ML by engineering features using scaling, attribute decomposition and attribute aggregation.

**Training Dataset** – In ML, a common task is the study and construction of algorithms that can learn from and make predictions on data. The model is trained on the training dataset using a supervised learning method. The training dataset is mainly used to allow machine to understand its operating environment.

**Model ­**– After defining the variables that we are going to use for the analysis, it will be time to use neural designer that will build predictive model for the sales of the stores

**Test Dataset** – The dataset used to provide an unbiased evaluation of a final model fit on the training dataset. Before using the model to forecast the sales, the last step will be to determine its predictive power on an independent set of data that has not been used before for the training. It is used to evaluate the accuracy at which the model can properly fit unseen data.

**Predicted Result** – The final results of the sales data about the future from the historical data.

# **8.3. The description of data**

In this study the following tables contains:

* Table 1: Anonymised information about the 45 stores, indicating the type and the size of the store. The table contain 45 rows and 3 columns.
* Table 2: Additional data related to the store, department, and regional activity for the given dates. The table contains 8190 rows and 12 columns.
* Table 3: Historical sales data, which covers from 2010 to 2012 within this tab. The table contains 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 |
|  |  |  |

* 1. **Data Splitting and Testing**

The use of ML algorithms improves the intelligence of the machine without manual intervention. “ML is used to optimize the performance criterion using sample data or past experience”. Three ML algorithms can be applied to prediction, Generalized Linear Model (GLM), Decision Tree (DT) and Gradient Boost Tree (GBT) will be implemented on the training dataset and models will be tested for the performance. Based on the prediction accuracy the best algorithm will be chosen for the prediction.

Firstly, we will use the descriptive analytics, which is the study of sales distribution, data visualization with different pair plots. It will be helpful 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. We will consider the supervised Machine-Learning approach using sales historical time series. For the case study, we will use Random Forest algorithm (Breiman, 2001). The accuracy on the testing set is an important indicator for choosing an optimal number of iterations of machine-learning algorithms. The effect of machine-learning generalization will enables us to make prediction in case of the very small number of historical sales data, 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.

We will have to split a historical data set on the training set and testing set by using period splitting, so the training data will lie in the first time period and the testing set in the next one. In this study we will use python to split our data into two subsets: training data and testing data, and fit our model on the train data in order to make predictions on the test data. When fitting our model one of the two things might happen which is either over-fit or under-fit our model. After selecting the best model + tuning, we will make predictions using the 20% test data to evaluate if it is performing well on unseen data. This will be applied to avoid model over-fitting. Over-fitting means, the model was trained “too well” and fit too closely to the dataset. This usually happens when the model is too complex. On the other hand, under-fitting refers to the model’s inability to fit in the training data and therefore misses the trends in the data. It also means the model cannot be generalized to new data.

Now we will use the Pareto Principle also called the 80/20 rule. The general point is that, in most cases, 80% of effects come from 20% of causes. Hence we will split our data into train\_test\_split using the sklearn library. The test\_size= 0.2 (20%) and the train\_size = 0.8 (80%) which data will be split by Pareto Principle (80/20) in a randomised way to test for over-fitting and under-fitting then fit the model on the training data trying to predict the test data.

# **Software tools**

* + 1. **Python**

Python will be used for developing predictive model and data analyses. The advantages of using python environment is its flexibility, low learning curve, well supported and documented. It has heavily been used in academic and industrial circles. Python has plenty of useful analytic libraries that will be added in the process of developing predictive models. Python will also be used for data cleaning to manage common issues such as inconsistent column names, missing data, and different data types and duplicate rows and etc.

We will install 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 Itel(IR) Core i5-6300U. We will also import necessary python libraries for data manipulation and data analysis. Database querying tool import data manipulation tools such as pandas and numpy. The imported libraries will help create data statistics and clearly visualised data to gain better insight and clearly explain the trends. Seaborn and matplotlib.pyplot libraries will be used for data visualization.

# **Scientific contribution**

Predictive analytics not only helps in creating practically useful models but also plays an important role in building new theory for further study and research. The use of available data to extract inferences and predictions by using predictive analytics has grown in the industry from being a small department in large companies to being an active component in most mid to large sized organizations. Sales forecasting can have a crucial impact on the success and performance of companies. Sales forecasting is a crucial part of the financial planning of a business, with an accurate forecast in hand we can plan for the future. Sales forecasts are also an important part of starting a new business with an accurate business plan and better decision making.

# **References**

# 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.

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.

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.

Cheriyan, S., 2018. *Intelligent-Sales-Prediction-Using-Machine-Learning-Techniques.* s.l., www.researchgate.net.

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

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., n.d. Sales Forecasting for Retail Chains.

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

Kharratzadeh, 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.

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].

Pavlyshenko, B., 2018. *Machine-Learning Models for Sales Time Series Forecasting.* Lviv, Ukraine.

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].

Thoben, K.-D., 2015. A survey on retail sales forecasting and prediction in fashion markets. In: *Systems Science & Control Engineering.* Mrtimer House, 37-41 Mortimer Street, London WIT 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.