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

BY

SINETHEMBA NONGQOTO

STUDENT NAME

201636304

STUDENT NUMBER

**RESEARCH METHODOLOGY**

SUPERVISOR: JOSEPH TLOUYAMMA

MONTH, YEAR MAY 2020

**RESEARCH METHODOLOGY**

# **CHAPTER 3: METHODOLOGY**

* 1. **Introduction**

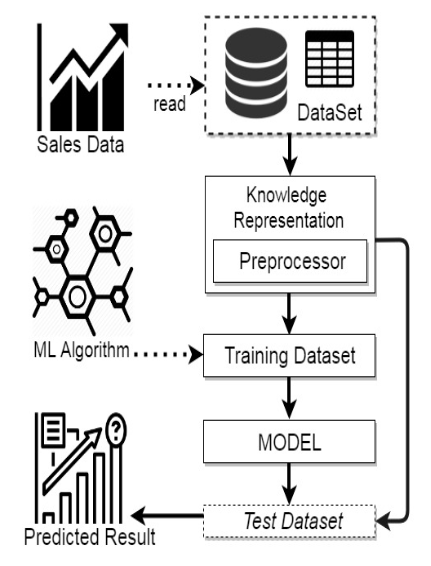
In this section we will specify the procedures or techniques used to identify, select, process and analyse information about retail data analytics. Data collection enables a person or organization to answer relevant questions, evaluate outcomes and make predictions about future probabilities and trends. 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, Kaggle: Retail Data Analytics, 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

After data pre-processing, in order to clearly understand the nature of our data, an exploratory analysis will be conducted (W. Huang, 2015). The exploratory analysis consists of the steps as shown in the Figure 1.



*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.

* + 1. **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 |
|  |  |  |

In our study since the data provided is in multiple files, with some of the columns present in more than one file. We will join and merge DataFrames since Joining and merging DataFrames is the core process to start with data analysis and machine learning tasks. It is one of the toolkit which every Data Analyst or Data Scientist should master because in almost all the cases data comes from multiple sources source and files. We need to bring all the data in one place by merging and some sort of join logic and then we will start with our analysis.

* + 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.

**Exploring the results:** sales\_df\_train.shape, sales\_df\_test.shape,

**Output:** ((337256), (84314)), The train will have 337 256 values and then the test will have 84 314 values after we split data into train\_test\_split.

* + 1. **Software tools**

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

# References

Alpaydin, E. (2004). Introduction to Machine Learning (Adaptive Computation and Machine Learning). *The MIT Press* .

Breiman, L. (2001). Random forests. *Mach. Learn*, pp. 45, 5-32.

Cheriyan, S. (2018). Intelligent-Sales-Prediction-Using-Machine-Learning-Techniques. www.researchgate.net.

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

Singh, M. (2017). *Kaggle: Retail Data Analytics*. Retrieved July 22, 2020, from https://www.kaggle.com/manjeetsingh/retaildataset

Singh, M. (2017, September 01). *Kangle.* Retrieved from Retail Data Analytics: https://www.kaggle.com/manjeetsingh/retaildataset?select=stores+data-set.csv

W. Huang, Q. Z. (2015). A Novel Trigger Model for Sales Prediction with Data Techniques . *Data Science Journal, vol.14*.