Machine and Deep Learning Techniques Applied to Retail Telecommunication Data

by

Fariha Naz

M.Sc., University of Lethbridge, Alberta, Canada, 2015

BCIT, NED University of Engineering and Technology, Karachi, Pakistan, 2010

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Science

SIMON FRASER UNIVERSITY

Fall 2019

Approval

Name:	**Fariha Naz**
Degree:	**Master of Science (Computer Science)**
Title:	**Machine and Deep Learning Techniques Applied to Retail Telecommunication Data**
Examining Committee:	**Chair:** Binay Bhattacharya Professor
	Fred Popowich Senior Supervisor Professor
	Anoop Sarkar Supervisor Professor
	Jiannan Wang Examiner Assistant Professor
Date Defended:	**November 7, 2019**

Abstract

Telecommunication service providers have franchise dealers to sell their services and products to a wide range of customers. These franchise dealers are small-sized businesses working with a small financial budget and limited human resources for analyzing the performance of the business. There are numerous commercial business intelligence (BI) tools to monitor data and generate business insights. However, most of the retail entrepreneurs still use manual and/or simple techniques, having little time to dedicate to sophisticated BI tools.

In this work, we investigate machine and deep learning techniques to analyze some retail telecommunication business datasets. Specifically, we examine how nearest neighbor techniques, feed forward artificial neural networks, Bayesian classifiers, and support vector machines can be used with retail telecommunication data. As indicated by our initial results we have been able to achieve precision, recall, and f-measures of 95%, for the task of classification, demonstrating that we can categorize retail telecommunication data based on the gross profit. We also developed a variant of recurrent neural networks (RNN), specifically Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) deep neural network models. Based on our initial results, we are able to acquire the root mean square error of 191 (training) and 281 (testing) from developed univariate models. A feed forward artificial neural network is applied to perform binary classification where we obtain an accuracy of 85% when categorizing the dataset based on the product type.

Keywords: deep learning, retail sales, neural networks, telecommunication, LSTM, time series forecasting, machine learning, sales, Android, iPhone, iOS, profitability, retail, feed forward artificial neural networks, bayesian classifier, nearest neighbor, support vector machines, supervised learning.

Dedication

This work is dedicated to myself, my son Muhammad Omer Ali, my husband Ali K., my courageous mother R. Syed, my siblings M. Rehman and M. Nauman.

Acknowledgements

I would like to thank Fred Popowich for giving me the opportunity and for his continuous support throughout the degree program.

I would like to extend my thanks to AH. I am grateful for receiving the understanding and continuous support from my husband Ali K. and brother M. Rehman, throughout the degree program.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the advent of mobile devices and mobile apps there are a number of companies that provide telecommunication services to customers. Telecommunication service providers, including TELUS, Bell, Fido, and Freedom Mobile have franchise dealers to sell their services and products to a wide-range of customers. The franchise dealers are entrepreneurs responsible for small-sized or medium-sized business. We are proposing to investigate the use of various machine and deep learning techniques towards the analysis of a retail telecommunication business dataset.

1.1 Motivation

Traditionally, various retail businesses use domain-specific POS (point of sale) systems to manage inventory and sales [14]. Example systems include Vend POS (by Vend), Shopify POS, Oracle Hospitality POS, QuickBooks (by Intuit), RQ by iQmetrix and many others. These telecommunication service providers have franchise dealers that sell their products to customers across different demographics. The small-sized business (franchise dealers) usually need to analyze the key performance indicators (KPI's) of the business. A Key Performance Indicator (KPI) is a metric used to measure performance, which is usually dependent on the business need. For example, one retail store might want to manage their inventory better, so they would use KPIs like inventory to sales ratios. Machine learning techniques are employed to measure the performance of a business in different domains. Examples are sales forecasting [57, 38], investigation of the loyalty of customers

[15], prediction of weekly demand for grocery items in a supermarket [51], and developing a forecasting model for airline business [22].

In this research work, we are interested in applying different techniques on datasets that are different in sizes as well as the number of features. With the presence of 400 plus rows and eleven features we attempted to apply nearest neighbor (K*) [13], support vector machines (SMO) [43], and naïve bayes[29] to a retail telecommunication business dataset. The above three supervised learning methods are used to solve various problems in different domains, including evaluation of machine translation quality [6], the analysis of breast cancer data [48], and categorizing computer programs based on gender [41].

Artificial neural networks are more favorable towards the forecasting and analyzing retail sales [2] because they are able to capture the dynamic nonlinear trend, seasonal patterns and interactions between them. Multilayer Perceptrons are applied in the domain of e-commerce, fashion, retail, and aviation. Various neural networks algorithms are employed to perform sales forecasting [57, 38], to understand the loyalty of customers towards the instant coffee business [15], to predict the future values of time series that consist of the weekly demand of grocery items in a supermarket [51], and airline passenger forecasting [22]. We propose to categorize the dataset using multilayer perceptron, which contain 29000 plus instances and eleven features.

Recently, deep learning is employed in the retail ecosystem to protect the privacy and security of consumers [23]. However, the big retailers including Best Buy, Walmart, Target and Shaw are slow to adapt state-of-the-art technologies like artificial intelligence and deep learning [19]. With the presence of large amounts of data and the advancement in the artificial intelligence including deep neural networks it is possible to analyze the dataset to forecast the performance of the retail business in the telecommunications domain. For this reason, we propose to investigate the machine and deep learning techniques towards the analysis of financial dataset related to Telecommunications.

Many researchers explored deep learning methods towards the problem of forecasting sales in the retail business and the energy usage [53]. Deep neural networks are used in the retail industry [53] to perform the demand forecasting, to forecast sales in a business [37], and to forecast the price of agricultural products [54]. Therefore, we applied deep neural networks mainly Long Short-Term Memory Network (LSTM) and Bidirectional Long Short-Term Memory Network (BiLSTM), a type of recurrent neural networks to analyze the univariate time series. The retail telecommunication business dataset contains more than 74,000 rows. Our approach will fill the gap which occurs due to the lack of information about the performance of the business in terms of categorization of products and profits.

Our models can be used by retail entrepreneurs in the telecommunication domain to forecast the performance of their business on the quarterly as well as yearly basis. Our models are integrated into the telecommunication business to leverage the retail company's historical financial, product and sales data. Our models will be used to predict profitability and forecasting product sales from multiple business locations.

1.2 Contributions

We investigate the supervised learning, neural networks and deep learning towards retail, product and time series datasets in telecommunication domain. The contributions of this research work are as follows:

- The research has investigated supervised learning techniques to categorize dataset on the basis of profit or loss.

- We develop neural network models using a variant of recurrent neural networks (RNN), specifically Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) deep neural networks.

- We analyze the dataset to develop artificial neural networks model in order to categorize dataset depending on the product type (or mobile devices).

- The developed models will be used by the retail telecommunication business with the future dataset to predict profitability, sales of popular mobile products as well as to forecasting the sales.

1.3 Thesis Organization

In Chapter 2, we shed light on the machine and deep learning techniques used in this work including nearest neighbor (K*), support vector machines, feed forward artificial neural networks (also referred as multilayer perceptron), naïve bayes, Long Short Term Memory (LSTM) and Bidirectional Long Short Term Memory (BiLSTM) deep neural networks. Moreover, we discuss various metrics employed to evaluate the performance of the developed models.

Chapter 3, we provide the open source retail dataset as well as the retail telecommunication business dataset used in this work. We also provide details about the various experiments that we performed.

In Chapter 4, the results of different experiments and technical implementations of the work are discussed.

In Chapter 5, we summarize the contribution of this research work and shed light on the possible future research directions.

Chapter 2

Related Work

In recent years, machine learning techniques became popular for the analysis of different kinds of data ranging from molecular data (genes or proteins) [42], text documents [4], twitter data[18], machine translation data [6], computer programs [41] and medical data [48]. Machine learning algorithms learn what trends are present in the given dataset which then can be used to make decisions for the analysis of future data. In the next sections, we will briefly cover machine and deep learning techniques used in this work including nearest neighbour (K*), support vector machines, feed forward artificial neural networks (also referred as multilayer perceptrons), naïve bayes, long short-term memory (LSTM) and bidirectional long short-term memory (BiLSTM) deep neural networks. Moreover, we discuss various metrics applied to evaluate the performance of the developed models. Parts of this chapter are taken from our work for [40], specifically details about nearest neighbour, support vector machines, and naïve bayes as mention in the next section.

2.1 Machine Learning

Machine learning algorithms are used to acquire knowledge and interesting information from diverse types of data. The algorithms are divided on the basis of how data can be dealt, for example, if there is the presence of class labels or not. There are mainly three types of machine learning methods [24, 55, 39] including supervised (or classification) learning, unsupervised (or clustering) learning, and semi-supervised learning.

Supervised learning is the process of determining a model that distinguishes data into different classes. As an example, numeric values like 0 and 1 may represent two different classes. Data samples that are present in the datasets may be associated with these class labels. The model is derived based on the analysis of a set of training data, that is, data for which class labels are known [24]. These models are called *classifiers*. The role of a classifier is to predict the categories of test data based on class labels. Classification can be performed using different machine learning algorithms including support vector machines, naïve bayes, and nearest neighbor [40] just to name a few. The procedure of classification is a two-step process that includes

1. learning (or training of a model), and
2. classification (or testing of a model).

Generally a dataset is represented by a set of features, and relevant features/attributes are usually unknown. The presence of irrelevant features sometimes degrades the predictive ability of the models. For this reason, a small set of features should be chosen [16, 5], consisting of features that are sufficient for learning and improving the quality of the concept description (model).

To further delve into the features, in this work, Information Gain, a statistical technique is applied to the given dataset. Information gain is calculated on the basis of the difference between the original information about the proportion of classes and the new information (as shown in Equation 2.1) that is obtained after the identification of the useful attribute. The information gain is calculated using the following formula [24, 55]:

$$\mathrm{InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute)} \tag{2.1}$$

For this reason, in this research work, we apply different machine learning techniques and deep neural networks. In order to perform the feature selection, an information gain is used to extract more relevant features. Long Short-Term Memory Network (LSTM), which is the variant of recurrent neural networks apply to develop univariate models for a retail telecommunication business dataset. In the next section, we will briefly discuss machine and deep learning methods which are employed in this thesis.

2.1.1 Support Vector Machines

The machine learning algorithm, support vector machines (SVM) is a kernel-based algorithm. The kernel function

$$K(x_i, x_j) = \varphi(x_i)^T \cdot \varphi(x_j) \tag{2.2}$$

calculates the dot product for the training vectors x_i in the possible high dimensional space [28]. In other words, training tuples (or data samples) of both linear (separable by a straight line) and nonlinear (not separable by a straight line) data can be mapped into the possible high dimensional space using kernel function (see Equation 2.2). In this work, we use the complex classifier technique [43] sequential minimal optimization method for training a support vector classifier.

2.1.2 Naïve Bayes

Naïve Bayes is a simple Bayesian classifier and predicts the probability of a given instance that belongs to a particular class. Naïve Bayes works on the assumption that each attribute that is associated with a given class is independent of other features. This is referred to as "class conditional independence" [30]. Naïve Bayes identifies the highest probability for a particular class label C_i, and chooses that label using the following formula [24, 55]:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{2.3}$$

In Equation 2.3, $P(C_i|X)$ is the probability that an instance X will belong to a particular class. The class with the highest probability is chosen [24]. $P(X|C_i)$ is the probability of X based on a specified class. $P(C_i)$ is the probability of instance that belongs to a particular class C_i. There are two class labels: profitable and non profitable. $P(X)$ is the probability of an instance X observed and is often constant.

2.1.3 Nearest Neighbor

K* is used to perform the supervised learning of a dataset [52, 34] and is considered suitable for continuous feature values. For example, dataset of medical reports is [48] analyzed using the K* algorithm. This is a nearest-neighbor method and implemented as a "nearest neighbor with generalized distance function" [55]. The K* classifier is based on the use of an entropic distance measure to provide prediction for future datasets (see Equation 2.4). Similar instances are identified using an entropy distance metric. The distance is calculated from the training sample. The new/test instance x is compared with the existing/training ones b_i using the distance metric [55]. Class labels are assigned to the new (test) instance on the basis of the closest k-nearest existing instances, b_i [25] as in the equation

$$K^*(b_i, x) = -\log P^*(b_i, x) \tag{2.4}$$

where $i \in \{1, 2, ...k\}$ and P is the probability of all possible paths from a training (b) to a test (x) instance.

2.1.4 Feed Forward and Deep Neural Networks

Neural networks learn from data without domain knowledge or feature engineering. The direction for the flow of data in a feed forward neural network is from input nodes to output nodes. There are three layers including input, hidden and output layers. The feedforward artificial neural network, as shown in Figure 2.1 [3], which is also called as the multilayer perceptron (MLP), employed in various domains to solve problems. The main feature of this algorithm [3] is that the information flows only in one direction and the output of a layer is determined from the previous layer. The example in Figure 2.1 shows the following:

- Four input features $X_1, \ldots, X_4$;
- A hidden layer which consists of five neurons $H_1, H_2, \ldots, H_5$;
- Two output units O_1 and O_2.

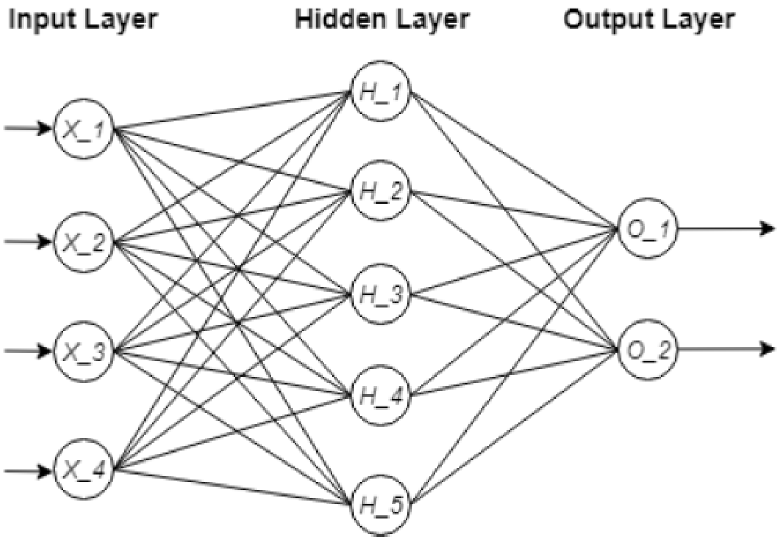

Figure 2.1: Multilayer Perceptron

A single matrix multiplication is used to compute the weighted sum for a hidden layer neuron. The hidden layer can be calculated using the following formula:

$$H = f(b + X * W) \tag{2.5}$$

Feed forward neural networks are commonly used when input data is of fixed size [3]. Neural networks learn from data without domain knowledge. For this reason, in the given empirical study we use a retail telecommunication dataset. Multilayer perceptrons are also used in other domains, for example, to evaluate the quality of machine translation [6] and the classification of breast cancer [48].

Recently, deep neural networks are studied in various domains to solve a broad range of problems [35]. For example, in the area of natural language processing (NLP), the two variants of the recurrent neural networks (RNN) including Long Short Term Memory (LSTM) and the Gated Recurrent Unit (GRU) showed improved performance in the machine translation and speech recognition [7]. Recurrent neural networks (RNNs), are a type of neural network, well-suited to time series data. In particular, Long Short-Term Memory (LSTM) [26] can capture long range dependencies and nonlinear dynamics. The structure of LSTM is similar to the standard recurrent neural network with the presence of one hidden layer. However, each node in a hidden layer is replaced by a memory cell. A memory cell is composed of an internal state (linear unit) and number of gates in order to control the flow of data. In the area of retail, LSTM is employed to perform sales forecasting for a real-world e-commerce database [8]. The convolution neural network (CNN) models are shown to have a positive impact on the performance in the area of computer vision [33]. In the health and medical field, LSTM is employed to perform supervised learning of the clinical time series [36].

Long short-term memory (LSTM) and Bidirectional LSTM (BiLSTM) [49] performed well across six benchmark datasets for the task of fine-grained sentiment tasks [9]. In the area of data mining, research work is carried out in the domain of retail or telecommunications business domain specifically, grocery stores [31]. Dual-stage attention-based recurrent neural network (DA-RNN) [45] is proposed to capture long-term temporal dependencies in order to develop multivariate time series neural network models based on the SML 2010 dataset and the NASDAQ 100 Stock.

2.2 Evaluation Methods

The predictive ability of a model (or a classifier) is determined by analyzing the confusion matrix and various evaluation metrics, including precision, recall, and f-measure. The values for these metrics range from 0% to 100%. The associated class labels of test data samples (or instances) are compared with the predictions produced by the model. If the results of the models are closer to 100%, than the model is considered useful to classify future (or unseen) data samples (or instances).

A confusion matrix demonstrates how well a model is able to classify the number of instances that are associated with various classes. The confusion matrix is a table of size ‘m x m’, as shown in Table 2.1, where m is the number of classes.

- true positives (TP) are positive tuples that are correctly predicted as being positive tuples;

- true negatives (TN) are negative tuples that are correctly predicted as being negative tuples;
- false positives (FP) are negative tuples that are misclassified as positive tuples; and
- false negatives (FN) are positive tuples that are incorrectly labeled as negative tuples.

Table 2.1: 2x2 Confusion Matrix.

Predicted Class				
Actual Class	Business	Profitable	Non Profitable	Total
	Profitable	TP	FN	P
	Non Profitable	FP	TN	N
	Total	P'	N'	P+N

2.2.1 Holdout Method and Cross Validation

The holdout method is used to divide the entire dataset into two independent sets which are partitioned using a fixed percentage of split [24]. One dataset is composed of the data samples/tuples that are used in the *learning* step, that is, in training and construction of the model. The other dataset consists of the tuples that are reserved for testing of the model (the *classification* step of the supervised learning [55, 24]). The test dataset must not be used during the learning step to reduce the chance of inaccurate predictions.

In the presence of scarce data the drawback of using the holdout method is that there is the possibility of overfitting or underfitting. Sometimes the knowledge from a given training dataset is not sufficient to accurately classify a test dataset. There is a high risk that random predictions made by the model will work and be 'learned'; that is, there is a risk of learning the incorrect information from the dataset. This can result when the model learns from the given training dataset, but on the test dataset only 50% of the data was classified correctly. This problem is called overfitting [16]. The opposite of this problem is underfitting in which the model is biased because it misses important information in terms of the relevant features and class labels.

In a situation where data is limited, the technique of cross-validation can be used to reduce the risk of overfitting [24, 55]. The reason that overfitting can occur is the generalization from the given training rows is not learned accurately [10, 28]. Therefore, the k-fold cross validation technique is commonly used to mitigate the risk of incorrect generalization of a dataset.

2.2.2 Metrics

There are different metrics that can be used to evaluate the performance of the models trained on two-class dataset including precision, recall, f-measure, and accuracy. For evaluated models developed using a numerical dataset the popular metric is root mean square error.

Accuracy is the measure of data samples that are classified correctly by the supervised learning model. Sometimes accuracy can be misleading when data is class-imbalanced [24]. Therefore, other metrics can be applied towards evaluating the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.6}$$

In some cases there tends to be an inverse relationship between precision and recall [24, 10]. F-measure is used to represent the combined measure of both precision and recall. F-measure represents the "harmonic mean of precision and recall" [24] as shown below [24, 55]:

$$\begin{gathered} F - measure = \frac{(2 * Recall * Precision)}{(Recall + Precision)}, or \\ F - measure = \frac{2.TP}{2.TP + FP + FN} \end{gathered} \tag{2.7}$$

For example, in experiment 1 (see chapter 4) with the data size of 409 rows we developed a support vector machine model and computed precision, recall and f-measure as shown below:

$$F - measure = \frac{(2 * 0.91 * 0.895)}{(0.91 + 0.895)},$$

$$F - measure = \frac{(1.6289)}{(1.805)}, \tag{2.8}$$

$$F - measure = 0.902$$

Precision is the measure of positive data samples that are classified as a specific class and actually belong to the specific class. It is the 'measure of exactness' [24] and can be computed as:

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

Recall is the measure of positive data samples that are correctly classified as a specific class label; however, there is no information about data instances that are mislabeled by the model. It is the 'measure of completeness' or 'true positive rate' [24], as shown in the following formula:

$$Recall = \frac{TP}{TP + FN}, or$$

$$Recall = \frac{TP}{P} \tag{2.10}$$

The root-mean-squared error (RMSE) is a readily used error metric to measure the performance of regression models. RMSE measures the performance by computing the difference between predicted values and observed values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y - \hat{Y}_i)^2}{N}} \tag{2.11}$$

2.3 Summary

Based on the above literature review related to the machine and deep learning methods we found that the algorithms, including support vector machines, naïve bayes and K* nearest neighbor are employed to categorize the quality of machine translation data [6], twitter data [18], computer programs [41] and medical data [48]. The k-cross validation is applied and the well performed algorithms are naïve bayes and support vector machine with an accuracy of 97.2818%. K* classifies the tuples using more than one nearest neighbor by applying Euclidean distance. SVM algorithm finds the hyperplane which categorize the data instances on the basis of class labels. This method works well with a numerical dataset. Naïve bayes uses the joint probabilities and is commonly used for sentiment analysis and document classification. The metrics used to evaluate the performance of the models developed to categorize the twitter data [18] as well as breast cancer dataset [48] include precision, recall and f-measure.

LSTM is applied to perform the political analysis on twitter dataset [18]. Long short-term memory (LSTM) also performed well across six benchmark datasets for the task of fine-grained sentiment tasks [9]. The dimensions of textual datasets are very high starting from 50 to 600. In the domain of retail or telecommunication business datasets, specifically, the grocery store [31] LSTM method is also being investigated. The experiments are performed by applying deep learning and logistic regression. The highest accuracy achieved is 86%. The variation in the accuracy of models occurred between 75% to 86% depending on the number of attributes. There are a higher number of attributes depending on the three different types of product categories, which are either 62 or 3312. In the health and medical field, LSTM is employed to perform supervised learning of a clinical time series [36]. The models are trained on 80% of the training and 10% of the testing data. There are 128 distinct labels indicating various health conditions, such as acute respiratory distress, congestive heart failure, seizures, renal failure, and sepsis. The best results are obtained with the use of 128 hidden units. The models are evaluated using the different metrics of area under the curve.

In short, the techniques including support vector machines, nearest neighbor (K*), multilayer perceptron, naïve bayes, Long Short Term Memory (LSTM) and Bidirectional Long Short Term Memory (BiLSTM) neural networks are applied in various domains and a wide selection of datasets. As mentioned in this chapter, we observe the results seems promising even with different domains and sets of data mainly retail, health, textual, twitter, computer programs, machine translation, and clinical dataset. With these insights in mind, we choose these methods as the basis to analyze a retail telecommunication business dataset. In the next chapter, we demonstrate the development of univariate models and categorization of retail telecommunication data based on the features, mainly gross profit and mobile product types.

Chapter 3

Methodology

In this chapter, we discuss the significant steps which we use to apply and evaluate various machine and deep learning techniques using a retail telecommunication business dataset. The goal of this work is to gain insights and develop supervised learning models that can classify data into profitable, non-profitable, android and iOS class labels. We perform several steps to employ machine learning methods on the retail business telecommunication dataset extracted from RQ Point of Sales. The first step is choosing the dataset and following the knowledge acquisition (KA) activity [32], which is the discussion with the domain experts to build a data corpus. The second step is the content selection to extract useful and informative features from the dataset. Lastly, we perform different experiments to develop various supervised learning models by analyzing a number of attributes, by applying the binary classification, and using deep learning methods on our retail telecommunication business dataset as well as employing testing protocols. Parts of this chapter are taken from our work for [40], specifically experiments 1 and 2.

3.1 Data Description

Recently, researchers analyzed public dataset which is available at the UCI Machine Learning Repository [46, 17]. For example, there is a daily demand forecasting orders dataset [20], online retail dataset [12] and the SML 2010 dataset [56]. The daily demand dataset [20] is collected from a logistic company and composed of twelve attributes as shown in Table 3.1. The domain of this dataset is in retail, but not telecommunication and represents the various kinds of orders and banking information.

Table 3.1: Daily Demand Forecasting Orders Dataset [20].

No.	Attributes
1	Day of the Week
2	NonUrgent Order
3	Urgent Order
4	Order Type A
5	Order Type B
6	Order Type C
7	Fiscal Orders
8	Orders Controller
9	Banking Orders 1
10	Banking Orders 2
11	Banking Orders 3
12	Total Orders

SML 2010 [56] is a public dataset collected from a monitor system mounted in a house. The list of features are shown in Table 3.2. The dataset is used to perform an empirical study [45] using dual-stage attention-based recurrent neural network (DA-RNN). We observe there are different attributes not similar to the dataset used in this research work. The problem domains and attributes of open source datasets are different from the given retail telecommunication dataset. As shown in Table 3.1 and 3.2 there are different attributes which are not similar to the dataset used in this thesis research work.

Table 3.2: SML 2010 Dataset [56].

No.	Attributes
1	Date
2	Time
3	Indoor temperature 1
4	Indoor temperature 2
5	Weather forecast temperature
6	Carbon dioxide 1
7	Carbon dioxide 2
8	Relative humidity 1
9	Relative humidity 2
10	Lighting 1
11	Lighting 2
12	Rain
13	Sun dusk
14	Wind
15	Sun light west
16	Sun light in east
17	Sun light in south
18	Sun irradiance
20	Enthalpic motor
21	Enthalpic motor turbo

22	Outdoor temperature
23	Outdoor relative humidity
24	Day of the week

An online retail dataset [12] is a public dataset and is collected from a UK-based online retail store. The transactions are occurring between December 2010 to December 2011. There are eight attributes as shown in Table 3.3. The similarity between [12] and our retail telecommunication business dataset includes *UnitPrice* and *InvoiceDate*. The attribute *UnitPrice* is similar to *UnitCost* shown in Table 3.5, where as, *InvoiceDate* attribute is same as *SoldOn* in Table 3.6. However, the attribute *StockCode* does not represent the mobile operating systems as per our need.

Table 3.3: Online Retail Dataset [12].

No.	Attributes
1	InvoiceNo
2	StockCode
3	Description
4	Quantity
5	InvoiceDate
6	UnitPrice
7	CustomerID
8	Country

One point of sales (POS) system is commonly used by small-sized retail telecommunication businesses which is called as RQ Point of Sales [44]. The dataset we used in our experiments is collected from RQ. We did not perform feature engineering which is the addition of new feature in the given dataset. The data represent the main activities of the sales, inventory, as well as the finances. The dataset is collected over the period of different number of years. The dataset used in this work is composed of different types as described below:

1. Financial Dataset: The dataset is composed of different attributes which represent information about the monthly profit made at different store locations over the period of three years. The size of dataset used in experiment 1 and 2 [40], as mentioned in section 3.2.1, is 400 rows and eleven features.

Table 3.4: Financial Dataset.

No.	Attributes	Data Types	Examples
1	Location	Categorical	L1, L2, . . . L13
2	Month	Categorical	Jan, Feb, . . . Dec
3	QuantitySold	Numerical	1039, 2037, . . .
4	QuantityRef	Numerical	-20, -34, . . .
5	NetQuantity	Numerical	448, . . .
6	TotalInvoiced	Numerical	55208.98, 2180.87, . . .
7	GrossSales	Numerical	30655.21, 40570.34, . . .
8	Cost	Numerical	46679.52, 62461.37, . . .
9	GrossProfit	Numerical	5455.68, 9719.50, . . .
10	HST	Numerical	4636.07, 6775.27, . . .
11	Total	Numerical	59845.05, 78956.14, . . .

12	Financial Status	Categorical	Profitable, Non Profitable

2. Product Dataset: The dataset is composed of different products sold on a daily basis at different periods in time over the duration of one year. We labelled the dataset into two classes which are based on the popular mobile operating systems: *Android* and *iOS*. The dataset analyzed in experiment 4, as discussed in section 3.2.2, contains 29000 rows and eleven features (shown in Table 3.5).

Table 3.5: Product Dataset.

No.	Attributes	Data Types	Examples
1	Quantity	Numerical	1, -1
2	UnitCost	Numerical	34.00, 79.00, 834.00…
3	OriginalUnitPrice	Numerical	149.0, 649.0, …
4	ListPrice	Numerical	49.00, 79.00, 849.00…
5	SoldFor	Numerical	49.00, 49.00, 749.00…
6	AdjustedUnitPrice	Numerical	49.00, 0, 109…
7	Total Sales	Numerical	49.00, 49.00, 749.00…
8	Total Discount	Numerical	0.00, 30.00, 100.00, …
9	HST	Numerical	82.81, 82.81, 1265.81…
10	TotalCost	Numerical	34.00, 284.00, 134.00…
11	GrossProfit	Numerical	15, 185, …
10	Product Type	Categorical	Android, iOS

3. Sales Time Series Dataset: The data points are collected over the period of three years. The dataset is composed of 74000 rows and two features which are shown in Table 3.6. The sales time series dataset is used in the experiment mentioned in section 3.2.2. We used the dataset to develop univariate time series models using deep neural networks. The variant feature on which we are performing univariate analysis is 'GrossProfit' with respect to the time series feature which is 'SoldOn' with a data type of DateTime(DaT).

Table 3.6: Sales Time Series Dataset.

No.	Attributes	DT	Examples
1	SoldOn	DateTime	2016-01-18 3:17 PM, ...
2	GrossProfit	Numerical	15.00, 18.75, -30.00 ...

To perform various experiments we have to anonymize, clean, and preprocess the extracted dataset in order to develop the retail corpus. We remove currency symbols, null values, sensitive information, customers information and employees information. We replace the original sensitive details with generic words. For example, in Table 3.4 we replace the original address of the stores with generic location names (L1, L2 ...). Using a knowledge acquisition (KA) technique [32], we remove irrelevant attributes such as customer name in order to create a retail business data corpus. As shown in Table 3.4, the datatypes are categorical and numerical.

To perform supervised learning the class labels are assigned. In our retail corpus, we use pre-defined class labels which are *profitable* and *non-profitable*. If the value of gross profit attribute contains zero or a negative value for a given instance, then the data instance is labelled as *non-profitable*. The positive value of gross profit is assigned a *profitable* class label as listed in Table 3.4. The class labels used to categorize the dataset based on the product type (as shown in Table 3.5) represents the main two types of operating systems, that is, *Android* and *iOS*.

3.2 Experimental Work

In this section, we shed light on various experiments performed using various techniques as well as different datasets and their sizes. We investigate machine learning and deep learning techniques towards the analysis of the retail telecommunication dataset. The use of domain-specific POS (point of sale) system is prevalent in retail businesses to store data related to the sales, inventory, finances, and employees. The RQ Point of Sales software [44] is commonly used by small-sized telecommunication retail businesses. The data is extracted from the retail software, representing the main activities of the product and sales, which are accumulated over the period of three years or less. In this thesis, the first step is the preprocessing of the dataset. The data corpus is shown in Table 3.4, 3.5, and 3.6.

In the next section, we will briefly discuss the experiments performed in this thesis work. In the first two experiments we employ three supervised learning methods including nearest neighbor, naïve bayes and support vector machines. In the third experiment, we develop univariate models using deep neural networks mainly the variant of RNN, which is Long Short-Term Memory Network (LSTM). Lastly, we perform binary classification based on the product type using a multilayer perceptron.

3.2.1 Experiment 1 and 2

In the experiment 1, as mention in [40], we perform supervised learning of the retail telecommunication business dataset, which is collected each month at 13 different locations for the period of three years. We employ three supervised learning algorithms: nearest neighbor (K*) [13], support vector machines [43] and naïve bayes [29]. Each of the algorithms are employed in this work by using WEKA [21]. The dataset has twelve attributes (including the class labels) as shown in Table 3.4. The pre-defined class labels are *profitable* and *non-profitable*. There are approximately 400 rows present which are associated with these class labels. As shown in Figure 3.1, the total number of data instances labelled as *non-profitable* are 165, where as, the total number of 244 data rows are labelled as *profitable*. Therefore, we can say that our dataset is not class imbalanced.

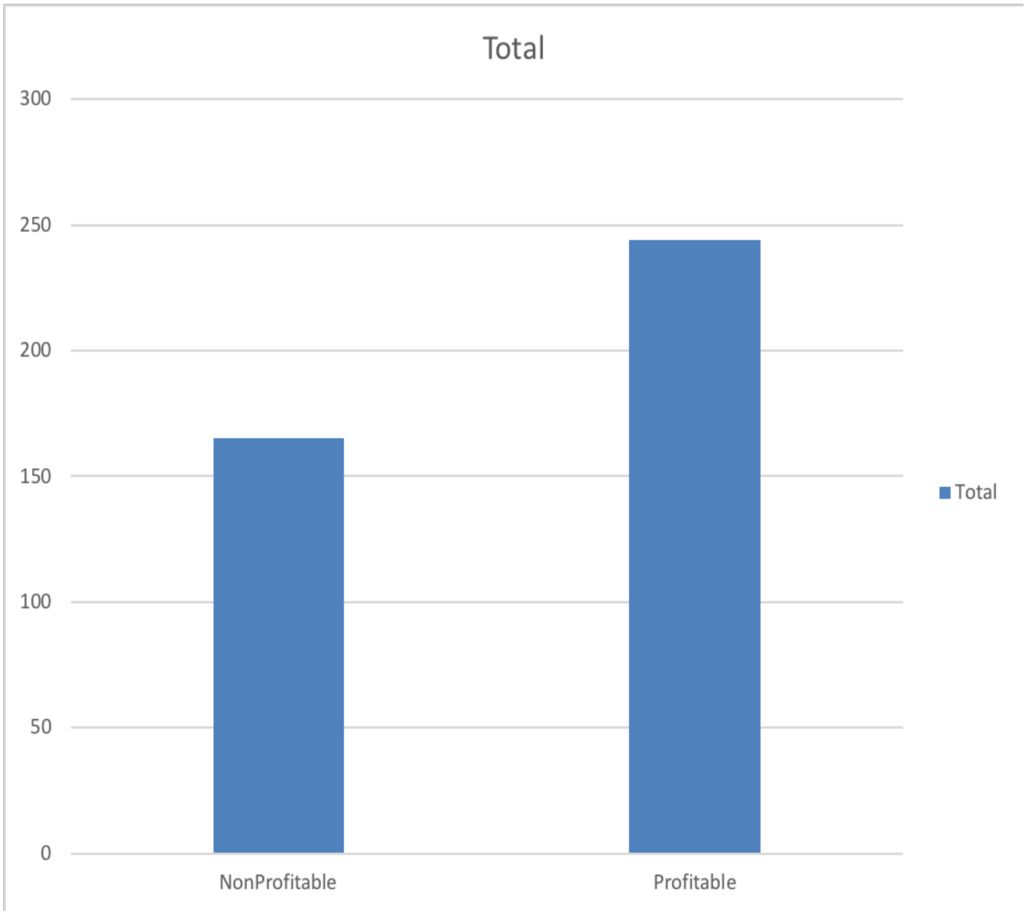

Figure 3.1: Class Distribution

In experiment 2, we applied single-attribute evaluator on the given dataset, which evaluates features on the basis of Information Gain. Feature selection is performed on the dataset to identify relevant features that are sufficient for learning and improving the quality [16, 5] of the concept description (model). Information gain is calculated on the basis of the difference between the original information and the new information that is obtained after the identification of the most useful attribute (as shown in Equation 2.1). The attribute with the highest information gain uniquely identifies each instance (shown in Table 3.7) which can either be retained or discarded to improve the performance of the predictive models. In short, in this experiment we performed content planning [50] which is the identification and selection of the important input features.

Table 3.7: TopRanked Attributes.

Attributes	Score
GrossProfit	0.9729
Month	0.4952
Location	0.1839
Cost	0.1401
HST	0.0969
TotalInvoiced	0.0826
Total	0.0788
GrossSales	0.0751

3.2.2 Experiment 3 and 4

In these experiments, our goal is two fold. The first goal is to develop univariate models using a deep neural network, specifically variants of RNNs known as Long Short-Term Memory Network (LSTM) and Bidirectional LSTM (BiLSTM). The second goal is to employ feed forward artificial neural networks to develop supervised learning models on the basis of two-class labels. We analyze the dataset collected from the domain of retail and telecommunication.

In experiment 3, we perform time series analysis by applying a deep learning method to train models for the dataset which is composed of 74,000 rows and two features. As shown in Figure 3.2, we can see that gross profit seems an increase over the period of year. Therefore, we apply deep neural networks specifically LSTM and BiLSTM recurrent neural networks [36, 9] to develop univariate models for the sales time series dataset as shown in Table 3.6. The variant feature is named as 'GrossProfit' which changes based on the time. The single column of gross profit is divided into two-column dataset: the first column contain present transaction (t) count and the second column contain the next (t+1) transaction count which will be predicted. The number of previous time steps are used as input variables to predict the next time period. To illustrate, X is the sales at a given time (t) and Y is the number of sales at the next time (t+1). We use 'MinMaxScaler' to transform the dataset and to rescale the data to the range of 0 to 1, which is also called as normalizing. In this work, we normalize the dataset using the MinMaxScaler preprocessing class. We split the dataset into two parts: a training and a test set using hold out method. The 70% of observations are used to train the models leaving the remaining 30% for testing the models.

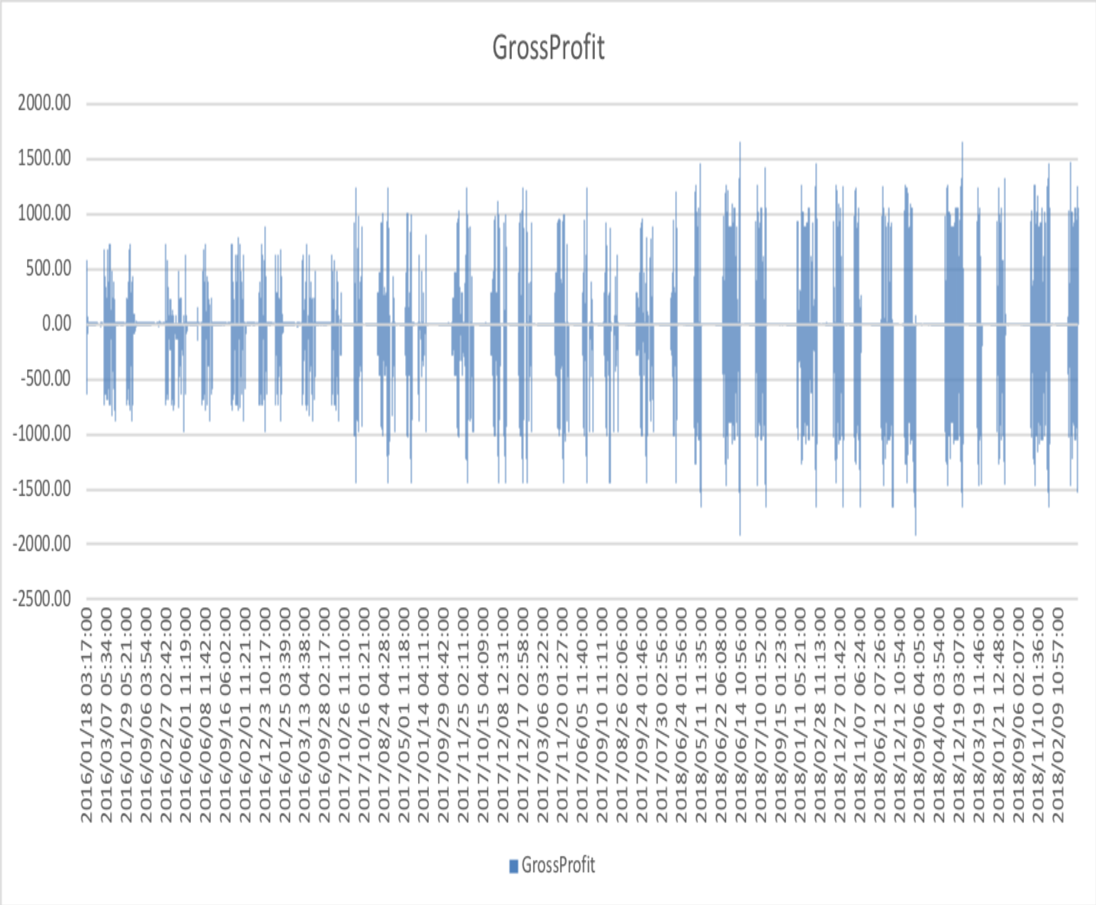

Figure 3.2: Gross Profit Over Time

In experiment 4, we used feed forward artificial neural networks (also known as multilayer perceptrons) to perform binary classification based on the mobile phones which is named as the product dataset, as shown in Table 3.5. The dataset is composed of numerical attributes. The size of the dataset is approximately 29000 rows and it consists of eleven features. Each instance is categorized based on the mobile device product type. There are two class labels which represent the main types of products, that is, *Android* and *iOS*. We perform *10-fold* cross validation to develop a classifier and to evaluate the performance of multilayer perceptrons model we used the accuracy metric. The measure of accuracy shows how well a classifier (or model) is able to correctly predict. The supervised learning model is trained on our product dataset. As shown in Table 4.4, we apply multilayer perceptrons to perform binary classification based on the product categories. In the next chapter, we will discuss and presents the results collected from our experiments.

Chapter 4

Results and Analyses

In this work, machine and deep learning models are developed to categorize and to analyze retail telecommunication business datasets. In the first two experiments (as described in section 3.2.1) we use financial dataset (Table 3.4)to develop three machine learning models using nearest neighbor (K*) [13], support vector machines [43] and naïve bayes [29]. The measure of accuracy is not used in these experiments because of the presence of sparse dataset. For this reason we used different metrics to evaluate the performance of the models trained on two-class dataset including precision, recall, and f-measure. Mac Book Pro is used to run experiments 1 and 2. The system running High Sierra 10.13.6, with an Intel Core i5 processor with the speed of 2.3 GHz, and 8 GB of memory. For both experiments we use open source tool WEKA [55, 27]. WEKA consists of various machine learning algorithms implemented by Witten and Frank. The code is written in the Java programming language to develop various machine learning models and to perform feature selection.

With the presence of a dataset which consists of 400 rows and named as the financial dataset (as shown in Table 3.4), we evaluate the performance of our three supervised learning models in experiment 1 and 2 (see section 3.2) by applying the k-fold cross validation. We set k = 10 and performed *10-fold* cross validation [40]. There are other cross validation techniques such as leave one out or applying k-fold with different number of folds. However, the *10-fold* cross validation is readily used by other researchers [24, 55] in different domain areas, especially when data is limited to reduce the risk of overfitting.

There are different metrics used to evaluate the performance of models trained on two-class dataset including precision, recall, f-measure, and accuracy [24, 55]. Sometimes accuracy can be misleading in the presence of a small number of data instances. Therefore, in experiment 1 and 2, other metrics are considered towards evaluating the three learning models. In Tables 4.1 and 4.2, we list the performance of trainable models and the evaluation metrics, including precision (PR), recall (RE), f-measure (FM). To further delve into the actual number of data instances that are correctly classified we use two additional columns true positives (TP) and true negatives (TN) which are collected from the confusion matrix. The column TP represents how many rows are correctly classified as profitable and TN demonstrates the number of instances labelled correctly as non-profitable by the supervised learning models.

Table 4.1: Machine Learning Models Performance

Models	**Precision (%)**	**Recall (%)**	**F-measure (%)**	**TP**	**TN**
K*	95.5	94.7	95.1	231	154
Naïve Bayes	91.6	94.3	92.9	230	144
Support Vector Machines	89.5	91.0	90.2	222	139

We perform a comparative analysis of various learning models on the basis of f-measure due to the small size of dataset to address concerns where precision and recall may not provide an accurate assessment of a model's predictive capabilities. As shown in Table 4.1, we achieved the lowest f-measure of 90.2% with the support vector machine because of the small number of correctly classified profitable data instances. K* model achieved the highest f-measure of 95.1%. This means that the classification models developed in this study may be used to categorize the retail telecommunication data instances. We observed in experiment 2 (see section 3.2.1) that after the selection of eight attributes the performance of our support vector machine learning model on the basis of f-measure slightly increased. As shown in Table 4.2, the performance of K* slightly decreased. We can say that there is no effect on the performance of naïve bayes, hence, we can conclude that the application of feature selection on the retail telecommunication dataset has little or no impact on the performance of the models as well as on the f-measure values of supervised learning models.

Table 4.2: Machine Learning Models Performance After Feature Engineering

Models	**Precision (%)**	**Recall (%)**	**F-measure (%)**	**TP**	**TN**
K*	94.3	94.7	94.5	231	151
Naïve Bayes	91.6	94.3	92.9	230	144
Support Vector Machines	89.9	91.0	90.4	222	140

In experiment 3 and 4 (as discussed in section 3.2.2) we build various learning models including feed forward neural network, long short-term memory networks, and bidirectional long short-term memory networks. The experiments are performed using state-of-the-art advanced research computing (ARC) systems, storage and software. Computations are performed on the heterogeneous cluster 'graham' [11], located at the University of Waterloo, from 'Simon Fraser University', managed by West Grid and Compute Canada. We develop machine learning models using the open source tool **Tensorflow** [1]. To run experiments, it takes a minimum of 15 mins to a maximum of four hours depending on the values of parameters tweak to achieve the best results. We perform various experiments by tweaking different parameters specifically number of epochs, optimizers, batch size, number of neurons, learning rate, drop out rate, activations, percentage of training and testing data. The programming language use for this purpose is Python3.

LSTM and BiLSTM algorithms are applied, on the sales time series dataset (as shown in Table 3.6) with the size of 74000 rows, to develop regression models for retail telecommunication business dataset in experiment 3. The root-mean-squared error (RMSE) is a readily used error metric to measure the performance of regression models. The predictor variable is a real number and the quality of predicted values is computed using the square of the error, taking the mean across all test instances and then calculating the square root. The score calculation indicates how good or bad a model is performing. Good prediction can be shown in the case of the lower RMSE values and the bad performance of the model is represented by the higher RMSE values. The results discussed in Table 4.3 are collected by applying Long Short-Term Memory Networks (LSTMs) to train and develop the univariate models. In this experiment we used the holdout method (as discussed in section 2.2.1) 70% of the dataset is used for training the model, whereas, 30% of the dataset is considered as test dataset.

Table 4.3: Long Short-Term Memory (LSTM) and Bidirectional LSTM Performance.

Models	EP	LR	TrRMSE	TsRMSE	OPT
LSTM	5	0.001	202.62	301.41	adam
LSTM	10	0.006	193.79	281.07	adam
LSTM+LRS	5	0.001	193.15	283.65	adam
LSTM+LRS	10	0.002	192.37	283.06	adam
LSTM+LRS	5	0.001	192.56	282.7	nadam
LSTM+LRS	20	0.002	**191.38**	**281.04**	nadam
BiLSTM+LRS	5	0.001	194.89	287.96	rmsprop
BiLSTM+LRS	10	0.004	194.41	287.02	rmsprop
BiLSTM+LRS	5	0.001	192.41	281.89	nadam
BiLSTM+LRS	20	0.002	**191.56**	**281.26**	nadam

In the Table 4.3, LSTM stands for Long Short-Term Memory and BiLSTM stands for bidirectional LSTM. RMSE shows the difference between the values predicted by a model and the actual observed values. We use 'MinMaxScaler' to transform the dataset, but the values of root mean square error are still high. There are different parameters tweaked to achieve the best performance of the LSTM and BiLSTM models. We tweak various parameters of deep learning models (shown in Table 4.3) including different types of activation methods, different learning rates (LR), various number of epochs (EP), learning rate scheduler (LRS), and multiple optimizers (OPT). Our initial results show that the bidirectional LSTM (BiLSTM) model is able to achieve the lowest root mean square error (RMSE) of 191.56 during the training and 281.26 in the testing phase. The LSTM model performed better than the bidirectional LSTM by achieving the lowest root mean square error of 191.38 (training) and 281.04 (testing). Hence, we observe the error rate reduces as well as the performance of neural network models improves after tweaking the number of epochs, learning rate of 0.002 and applying the NADAM optimizer.

Table 4.4: Feed Forward Neural Network Performance

Epochs	Batch	Accuracy (%)	Comments
5	5	83.36	-
5	5	**85.27**	Standardize
5	10	83.59	-
5	10	79.81	Dropout (0.5)

In experiment 4, we used feed forward artificial neural network (also known as multilayer perceptrons) to perform binary classification based on the product type. The dataset used is named as product dataset, as shown in Table 3.5. As discussed in the previous section, the experiments are performed at the heterogeneous cluster ‘graham’ [11], located at the University of Waterloo, from ‘Simon Fraser University’, managed by West Grid and Compute Canada. **Tensorflow** [1], an open source tool, is used to run the experiments and to develop feed forward neural network models. The programming language use for this purpose is Python3. In this experiment, the dataset is composed of numerical attributes and two class labels in order to perform the binary classification. There are two class labels which represent the main types of products, that is, *Android* and *iOS*. The size of the dataset is composed of 29000 rows and eleven features, as shown in Table 3.5. Each instance is categorized based on the mobile device product type. We perform stratified *10-fold* cross validation to develop a classifier and to evaluate the performance of multilayer perceptrons model we use an accuracy metric.

The measure of accuracy shows how well a classifier (or model) is able to correctly predict, specifically when dataset has a large number of rows. As shown in Table 4.4, we employ multilayer perceptrons (or feed forward neural networks) to perform binary classification based on the product categories, that is, *Android* and *iOS*. In Table 4.4, we use different batch sizes, apply dropout regularization technique and standardization technique. The feature is standardized by removing the mean and scaling to unit variance. In the drop out technique, randomly selected neurons are ignored ('dropped-out') during the training phase. The contributions of those neurons are removed and weight updates are not applied. We observe the improvement in the accuracy to 85.27% when we apply the standardization technique to categorize the two-class dataset. We find out that a regularizer with the dropout value of 0.5 [9] reduced the accuracy of the trained model to 79.81% in order to deal with the overfitting (or incorrect generalization of the dataset during training phase).

Chapter 5

Conclusion

In this work, we examine the use of supervised learning on a retail telecommunication business dataset on a collection of class labels, specifically the profitable, non-profitable, Android and iOS. There are multiple experiments carried out in this research work including: binary classification to categorize dataset on the basis of profitability and sales of the product; and development of neural network models for univariate time series. To the best of our knowledge, this is the first work in the retail telecommunication domain and towards the development of learning models for the analysis of a retail business dataset. We conclude that for a retail small-sized dataset, it is possible to utilize machine and deep learning techniques to categorize the data instances and develop various supervised learning models.

We developed classification models to categorize the retail telecommunication business dataset on the basis of profitable and non-profitable class labels. We used a dataset which is composed of 400 rows and 11 attributes. Based on our initial results, shown in Table 4.1, we observed that nearest neighbor supervised learning model (K*) is able to achieve the highest f-measure of 95.1%. The K* nearest neighbor technique is used for continuous or numerical datasets, maybe this is one of the reasons the results are better. The method is a type of lazy learning where the function is approximated locally and computations are carried out during the classification. The lowest f-measure of 90.2% is obtained by a support vector machine. SMO is a complex support vector machine classifier that can be applied as a linear or nonlinear model to learn from the given dataset [47].

We performed feature selection to identify eight relevant features (Table 3.7) but the results are not drastically improved for the three supervised learning models developed (as described in section 3.2.1). We were hoping to see the increase in the performance of the models after the selection of eight attributes. However, we observed in Table 4.2 that the application of feature selection shows little to no improvements in the f-measure of the three supervised learning models.

To further extend this research work, we performed the empirical study of retail telecommunication business datasets using LSTMs, BiLSTM, and artificial feed forward neural networks. In this thesis, there are experiments using deep and neural networks including: binary classification using multilayer perceptrons to categorize datasets on the basis of mobile device product types; and development of univariate models for retail telecommunication business dataset. To develop the feed forward neural network classifier, the retail dataset is labelled into two classes which are based on the popular mobile operating system: *Android* and *iOS*. We used the variant feature on which we are performing univariate analysis is 'GrossProfit' with respect to the time series feature which is 'SoldOn' with datatype of DateTime(DaT) to develop univariate models using deep neural networks. To evaluate the performance of the trainable models we applied accuracy and root mean square error metrics.

Based on our initial results, as shown in section 3.2.2 and Table 4.2, we are able to categorize the retail telecommunication dataset with the highest accuracy of 85.27% in order to develop a feed forward neural network model. The results are improved after the application of a standardize technique which changes the values so that the standard deviation of the distribution becomes equal to one and outputs the normal distribution. We also notice the application of the drop out technique results in the reduction of the accuracy for the feed forward neural networks model.

We observed the results of the LSTM and BiLSTM models are improved by tweaking parameters including number of epochs, learning rate, learning rate scheduler, and optimizers. As listed in Table 4.3, the deep learning LSTM model is able to achieve the root mean square error of 191.38 on the training set and 281.04 on the test set. The results are not very promising and there is still a great potential to improve root mean square error and accuracy of product categorization in the retail telecommunication business domain.

5.1 Future Research Directions

In the future, we are interested to expand the proposed idea in different directions which are mentioned below:

- We are interested in the future to improve the rmse score by adding a more historical dataset, extracting relevant features, tweaking the number of neurons, batch sizes, epochs and other parameters.

- In this work, we applied six machine learning methods only. We observed that K*, feed forward neural network, and LSTM performed well. However, there are other methods including decision tree, random forest, and just to name a few. Hence, in the future, other machine learning methods will need to be investigated in order to find which algorithm is more suitable with a retail telecommunication business dataset.

- It would be interesting to investigate how to develop a framework for extracting data from different types of retail telecommunication software used in a business. Examples are Incomm and Moneris. These products are used for the payment processing.

- In the future, we are planning to extend the research for the analysis of multivariate time series using LSTM as well as dual-stage attention-based recurrent neural network (DA-RNN) for a retail telecommunication business dataset.

- The given dataset also contains 48 text summaries based on the retail telecommunications numerical dataset. The more interesting avenue is to develop a system for small-sized retail businesses in the telecommunication domain, which will generate textual reports in English instead of just graphs or numerical tables to track finances, sales, profit, and loss and other business activities. The report generator system will provide business activities and forecasting reports in natural (human) language to help small-sized retail entrepreneurs in decision making. Creating human-readable reports could also enable entrepreneurs to provide clear communication between stakeholders of the business as well as for non-experts.

Bibliography

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] I. Alon, M. Qi, and R. J. Sadowski. Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods. volume 8, pages 147–156, May 2001.

[3] A. Ambartsoumian. Applying self-attention neural networks for sentiment analysis classification and time-series regression tasks. Master's thesis, SIMON FRASER UNIVERSITY, 2018.

[4] S. Argamon, J. Goulain, R. Horton, and M. Olsen. Vive la différence! text mining gender difference in french literature @ONLINE. *Digital Humanities Quarterly*, 3(2), 2009. http://www.digitalhumanities.org/dhq/vol/3/2/000042/000042.html.

[5] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. Gender, genre, and writing style in formal written texts. *Text & Talk An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 23(3):321–346, 2006.

[6] B. R. Ayala and J. Chen. A machine learning approach to evaluating translation quality. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '17, pages 281–282, Piscataway, NJ, USA, 2017. IEEE Press.

[7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473, September 2014.

[8] K. Bandara, P. Shi, C. Bergmeir, H. Hewamalage, Q. Tran, and B. Seaman. Sales demand forecast in e-commerce using a long short-term memory neural network methodology. *CoRR*, abs/1901.04028, 2019.

[9] J. Barnes, R. Klinger, and S. Schulte im Walde. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. *CoRR*, abs/1709.04219, 2017.

[10] R. P. L. Buse and W. Weimer. Learning a metric for code readability. *IEEE Transactions on Software Engineering (TSE Special Issue on the ISSTA 2008 best papers)*, 36(4):546–558, 2010.

[11] Compute Canada. *Advanced research computing*. https://www.computecanada.ca/wp-content/uploads/2018/12/ComputeCanada-AR2018-EN.pdf.

[12] D. Chen, S. L. Sain, and K. Guo. Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. In *Journal of Database Marketing and Customer Strategy Management*, volume 19, pages 197–208, August 2012.

[13] J. G. Cleary and L. E. Trigg. K*: An instance-based learner using an entropic distance measure. In *12th International Conference on Machine Learning*, pages 108–114, 1995.

[14] C. Cumby, A. Fano, R. Ghani, and M. Krema. Predicting customer shopping lists from point-of-sale purchase data. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 402–409, New York, NY, USA, 2004. ACM.

[15] Y. Deliana and I. Rum. Understanding consumer loyalty using neural network. *Polish Journal of Management Studies*, 16:51–61, 12 2017.

[16] P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[17] D. Dua and C. Graff. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017. http://archive.ics.uci.edu/ml.

[18] T. Elghazaly, A. Mahmoud, and H. A. Hefny. Political sentiment analysis using twitter data. In *Proceedings of the International Conference on Internet of Things and Cloud Computing*, ICC '16, pages 11:1–11:5, New York, NY, USA, 2016. ACM.

[19] Daniel F. *Artificial Intelligence in Retail–10 Present and Future Use Cases @ONLINE*. https://emerj.com/ai-sector-overviews/artificial-intelligence-retail/.

[20] R. P. Ferreira, A. Martiniano, A. Ferreira, A. Ferreira, and R. J. Sassi. Study on daily demand forecasting orders using artificial neural network. 14:1519–1525, 2016.

[21] E. Frank, M. A. Hall, and I. H. Witten. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 4th edition, 2016.

[22] S. M. T. F. Ghomi and K. Forghani. Airline passenger forecasting using neural networks and box-jenkins. In *2016 12th International Conference on Industrial Engineering (ICIE)*, pages 10–13, Jan 2016.

[23] D. D. Gutierrez. *Reinventing the Retail Industry Through Machine and Deep Learning @ONLINE*. https://www.dellemc.com/content/dam/uwaem/production-design-assets/en-gb/solutions/assets/pdf/insideHPC-Report-Reinventing-the-Retail-Industry.pdf/.

[24] J. Han, M. Kamber, and J. Pei. *Data Mining Concepts and Techniques*. Elsevier and Morgan Kaufmann Publishers, 3rd edition, 2012.

[25] P. Hart. The condensed nearest neighbour rule. *IEEE Trans. Inf. Theory.*, 14(3):515––516, 1968.

[26] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation.*

[27] G. Holmes, A. Donkin, and I. H. Witten. Weka: A machine learning workbench. In *Proc. Australia and New Zealand Conf. Intelligent Information Systems*, Brisbane, Australia, 1994.

[28] C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification, 2010.

[29] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

[30] S. M. Kamruzzaman, F. Haider, and A. R. Hasan. Text classification using data mining. *Proc. International Conference on Information and Communication Technology in Management (ICTM-2005)*, May 2005.

[31] Y. Kaneko and K. Yada. A deep learning approach for the prediction of retail store sales. pages 531–537, December 2016.

[32] R. Kondadadi, B. Howald, and F. Schilder. A statistical nlg framework for aggregated planning and realization. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1406–1415, 2013.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[34] M. Kukreja, S. A. Johnston, and P. Stafford. Comparative study of classification algorithms for immunosignaturing data. *BMC Bioinformatics*, 2012.

[35] G. Lai, W. Chang, Y. Yang, and H. Liu. Modeling long- and short-term temporal patterns with deep neural networks. *CoRR*, abs/1703.07015, 2017.

[36] Z. C. Lipton, David C. Kale, and Randall C. Wetzel. Phenotyping of clinical time series with LSTM recurrent neural networks. *CoRR*, abs/1510.07641, 2015.

[37] M. Müller-Navarra, S. Lessmann, and S. Voß. Sales forecasting with partial recurrent neural networks: Empirical insights and benchmarking results. In *2015 48th Hawaii International Conference on System Sciences*, pages 1108–1116, January 2015.

[38] T. Choi . Hui N. Liu, S. Ren and S. Ng. Sales forecasting for fashion retailing service industry: A review. *Mathematical Problems in Engineering*, 2013.

[39] F. Naz. Do sociolinguistic variations exist in programming? Master's thesis, University of Lethbridge, Alberta, 2015.

[40] F. Naz and F. Popowich. Mining retail telecommunication data to predict profitability. In *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, Aug 2019.

[41] F. Naz and J. E. Rice. Sociolinguistics and programming. In *2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pages 74–79, Aug 2015.

[42] P. Pavlidis, I. Wapinski, and W. S. Noble. Support vector machine classification on the web. 20(4):586–587, 2004.

[43] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[44] Cell Phone Store POS and Retail Management Software @ONLINE. https://www.iqmetrix.com/products/rq.

[45] Y. Qin, D. Song, H. Cheng, W. Cheng, G. Jiang, and G. W. Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2627–2633. AAAI Press, 2017.

[46] UCI Machine Learning Repository. *Online Retail Dataset @ONLINE*. https://archive.ics.uci.edu/ml/datasets/Online+Retail#.

[47] J. S. Raikwal and K. Saxena. Performance evaluation of svm and k-nearest neighbor algorithm over medical data set. *International Journal of Computer Applications*, 50:35–39, 07 2012.

[48] H. Saoud, A. Ghadi, M. Ghailani, and B. A. Abdelhakim. Application of data mining classification algorithms for breast cancer diagnosis. In *Proceedings of the 3rd International Conference on Smart City Applications*, SCA '18, pages 84:1–84:7, New York, NY, USA, 2018. ACM.

[49] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. 45(11):2673––2681, 1997.

[50] P. K. Sowdaboina, S. Chakraborti, and S. Sripada. Learning to summarize time series data. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing 2014, pages 515–528, 2014.

[51] F. M. Thiesing and O. Vornberger. Sales forecasting using neural networks. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, volume 4, pages 2125–2128 vol.4, June 1997.

[52] S. Vijayarani and M. Muthulakshmi. Comparative analysis of bayes and lazy classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(8), August 2013.

[53] J. Wang, G. Q. Liu, and L. Liu. A selection of advanced technologies for demand forecasting in the retail industry. In *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, pages 317–320, March 2019.

[54] Y. Weng, X. Wang, J. Hua, H. Wang, M. Kang, and F. Wang. Forecasting horticultural products price using arima model and neural network based on a large-scale data set collected by web crawler. *IEEE Transactions on Computational Social Systems*, 6(3):547–553, June 2019.

[55] I. H. Witten and E. Frank. *Data Mining Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers, 2nd edition, 2005.

[56] F. Zamora-Martínez, P. Romeu, P. Botella-Rocamora, and J. Pardo. On-line learning of indoor temperature forecasting models towards energy efficiency, energy and buildings. 83:162–172, November 2014.

[57] K. Zhao and C. Wang. Sales forecast in e-commerce using convolutional neural network. *CoRR*, abs/1708.07946, 2017.