Sales Prediction Using Linear and KNN Regression

Shreya Kohli, Gracia Tabitha Godwin, and Siddhaling Urolagin

1 Introduction

Sales prediction plays a key role in building up a business. It is one of the most important parts of business intelligence. Sales prediction and forecasting give an insight into how a company should manage its workforce—labor, cash flow, and its resources. It is an evaluation tool that uses past and current sales data to predict future performance. Estimating future sales is an important part of the financial planning of any business. It enables companies to predict short-term and long-term performance. Accurate sales forecasts assist the companies in making informed choices which result in better supply chain management, an increase in profits, and better customer experiences. It is a crucial part of starting a new business as it helps in managing the available resources efficiently. Furthermore, with the help of data and gained insights, it becomes easier to understand consumer behavior. This gives way to the use of effective marketing techniques and can be used to selectively target the market. Some benefits of sales prediction for a business/company are that it can be used as a benchmark. It can also be used to help with planning. Demand and supply actions can be planned by looking at the forecasts.

Data analytics and predictive modeling are used for sales forecasting models. Predictive modeling is a process that uses information from data to determine the outcomes with data models. Many types of classifiers can be used to predict sales such as regression, K-nearest neighbor, decision trees, random forest, demand forecasting, classification methods, cluster analysis, and Bayesian classification. The

S. Kohli · G. T. Godwin (✉) · S. Urolagin
BITS Pilani, DIAC, Dubai, United Arab Emirates
e-mail: tabitha.godwin.tg@gmail.com

S. Kohli
e-mail: kohli.shreya27@gmail.com

S. Urolagin
e-mail: siddhaling@dubai.bits-pilani.ac.in

S. Patnaik et al. (eds.), *Advances in Machine Learning and Computational Intelligence*,
Algorithms for Intelligent Systems, https://doi.org/10.1007/978-981-15-5243-4_29

main classifier used in this paper is regression. The regression model equation might be as simple as $Y = a + bX$ in which case Y is your sales, the 'a' is the intercept, and the 'b' is the slope. With this model, we aim to correlate a variable that could be causing your sales to get better or worse. The two types of regression classifiers that are used in this paper are linear regression and KNN regression. The regression model has a few advantages. Firstly, the linear regression model is easy to interpret and understand due to its linearity. It can help businesses understand the relationship between various variables or factors affecting their profit. Furthermore, new data can be added easily which will not impact the accuracy of the KNN algorithm. It provides a powerful statistical method that can be used for data analysis.

This paper consists of a detailed literature review (Sect. 2), information regarding the dataset (Sect. 3). Then, there is a brief methodology (Sect. 4) where it tells how the data has been preprocessed and gives an insight into how feature selection has been done. Finally, we use linear regression and KNN regression models to train the dataset and extract the results. These results (Sect. 5) are then compared to conclude which regression analysis method is better for sales prediction on the Rossman dataset.

2 Literature Review

A large amount of data available in information databases becomes a waste until the useful information is extracted. Predictive analytics is known as the roof of advanced analytics—that is to predict future events. Predictive analytics is comprised of data collection and statistics, and deployment [1]. The rapid growth and advances of information technology enable data to be accumulated faster and in much larger quantities [2]. Predictive modeling is a combination of mathematical techniques that have in common the goal of finding a mathematical relationship between target with the purpose of estimating future values of those predictors and including them into the mathematical relationship to foretell future values of the target variable [3]. In every organization, the sales forecast is of utmost importance to help make decisions on every little and big detail, from budgets to spend to the labor required to profit, etc. Forecasting is a necessary task that helps ensure that an organization develops and plans successfully, but it is very much detested due to the amount of effort put into the process and how time-consuming the task can be. By feeding this data into such predictive analytic models, sales teams are now enabled with analytics-based insights and recommendations [4]. ML procedures have been gaining influence over time as interest in artificial intelligence has been growing [5]. Data mining means extracting information from the data, it means preparing data to gain the implied, prior unknown, potential and useful information, which can be represented as patterns [6].

Regression is an important analytical method used in teaching, economics, financing, etc., that computes the relationships between one dependent and one or more independent variable (s). The two primary kinds of regression are simple linear

regression and multiple linear regression. Simple linear regression applies one independent variable to predict the result of the dependent variable, and multiple linear regression relates two or more independent variables to foretell the result of the dependent variable [7]. Sales forecast is rather a regression problem than a time series problem. Study shows that the utilization of regression approaches can often give us more reliable outcomes opposed to time series techniques [8]. Sales forecasting is an important aspect of almost all businesses these days. Companies, nowadays, are attempting to expand such abilities of forecasting and prediction to get the upper edge on their competitors. For example, a good forecasting model gives information to manufacturer about the right amount of inventory, workforce, or labor required to satisfy the demand for the product [9]. The main goal of sales prediction is to analyze how internal and external factors can affect weekly sales in the future for Rossmann stores. Sales prediction is carried out using various machine learning and data science algorithms [10]. From prior literature, it can be noted that there has already been intensive research on three major uses of sales prediction. First is the Microsoft Time Series algorithm. It provides us with optimized regression algorithms for forecasting continuous real-time values [11]. Second is spatial data mining for retail sales forecasting [12]. Support vector regression (SVR), a technique is used in designing regression models to predict the expected turnover. Built from prior expert knowledge along with analytic knowledge discovered during data mining processes, this model provides us with accurate results. Finally, a novel trigger model for sales prediction with data mining techniques [13] that focuses on how to forecast sales with more accuracy and precision. It is now said that companies that can accurately forecast sales can successfully adjust future production levels, resource allocation, and marketing strategies to match the level of anticipated sales. A regression model is used to forecast or predict the value of the dependent variable—sales, based on various independent variables [14].

3 Sales Dataset

The dataset used in this paper is the Rossmann Store Sales available on Kaggle [15]. Rossmann is a German drugstore chain with 3466 stores under them (Table 1).

It can be observed from Fig. 1a, that the sales production decreased from 39.21% in 2013 to 23.66% in 2015. About 20% of sales were observed on the first day whereas only 0.51% was on the last day of the week. It can be concluded that most of the products were sold on a Monday and Friday. The maximum number of customers was observed to be 256 M in 2013 and decreased by 108–148 M in the year 2015. Figure 1b depicts the sales with respect to the number of customers visiting in the years 2013, 2014, and 2015.

It is observed from Fig. 2 that most of the sales occur when the store is running a promotion. Similarly, the number of customers visiting the store is relatively high during the promotion time. The fourth quarter showed an increase in average number of buyers.

Table 1 Attributes of Rossmann store sales

Attribute name	Description
Id	It shows the ID of the form (Store, Date)
Store	It is a distinctive number for the stores
Sales	Indicates revenue for a particular day
Customers	Gives the number of buyers during a particular day
Open	Shows if the store was open (=1) or closed (=0)
State holiday	a = public holiday, b = Easter holiday, c = Christmas, and 0 = None
School holiday	This shows if the (Store, Date) was affected by the closure of public school
Store type	Indicates the 4 stores that are a, b, c, and d
Assortment	Gives an assortment level: a = basic, b = extra, and c = extended
Competition distance	It is the distance to the closest competitor store in meters
Competition open since [Month/Year]	Shows the month as well as the year the closest competitor store was started
Promo	This shows if a store runs a promotion on a particular day
Promo2	Indicates if a promotion is continuous for some stores, 0—the store is not involved and 1—the store is participating
Promo2 Since [Year/Week]	This indicates the year and week during which the given store took part in Promo2
Promo interval	It gives the consecutive intervals Promo2 began

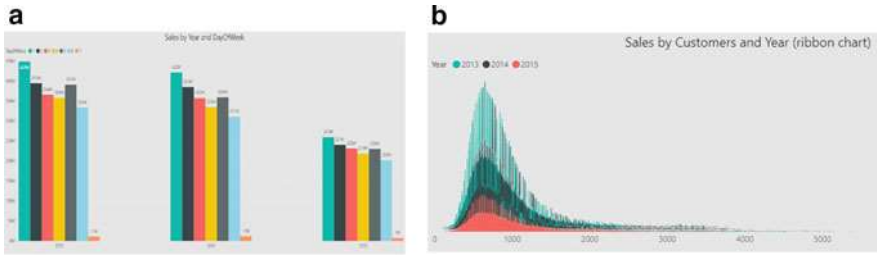

Fig. 1 Sales by (**a**) Year and days of the week. (**b**) Customers and year

Figure 3 depicts the distribution of sales and the distribution of customers. It can be observed that the standard deviation is smaller for the distribution of customers as the graph in Fig. 3a is narrower than the normal graph in Fig. 3b.

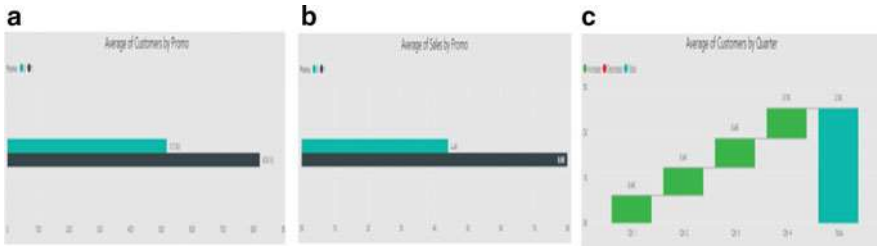

Fig. 2 (**a**) Average sales-promo, (**b**) average customers-promo, and (**c**) average customers-quarter

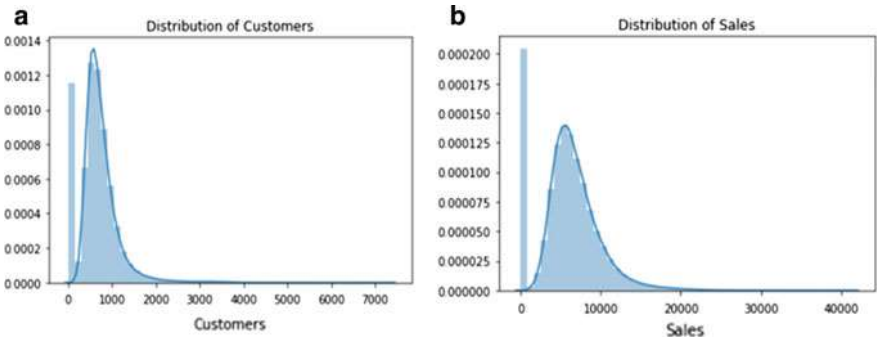

Fig. 3 (**a**) Distribution of customers and (**b**) distribution of sales

4 Methodology

The initial stage involves data collection and comprehension. This paper makes use of the Rossmann dataset from Kaggle. This data is then transformed into understandable form and the necessary features are selected. This is followed by predictive analysis using classifiers. Finally, the model is evaluated by applying various statistical methods (Fig. 4).

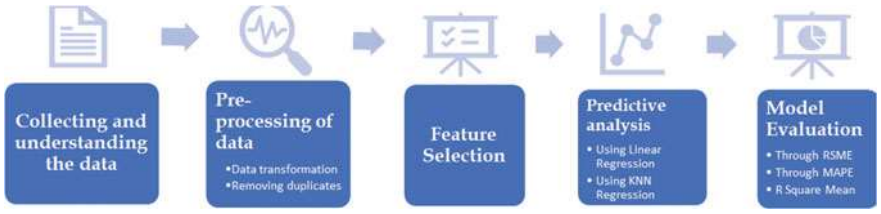

Fig. 4 Overall framework of sales prediction model used

4.1 Collecting and Understanding the Data

This is the process of gathering data and examining the dataset being used.

4.2 Data Preprocessing

Data preprocessing is a data mining method that includes converting raw data into an understandable form. This involves understanding the data, handling missing values, and removing duplicates.

4.3 Feature Selection

Feature selection is the process of finding the attributes or terms that are the most meaningful and extracting useful information from it.

4.4 Predictive Analysis

Linear Regression. Linear regression is a linear modeling approach to find the relationship between 1 or more independent variables (predictors) denoted as X and dependent variable (target) denoted at Y. Linear regression is all about finding the best fit line for the training as well as test data. The best fit line can be found by minimizing the distance between all data points and its distance to the regression line (by calculating the error (sum of squares error), we can find minimize distance), i.e., the distance between the points and the line should be minimum. This is done in a recursive method. The x value here varies between promo/customer/holiday, etc., and the y value here is sales.

K-Nearest Neighbor Regression. K-nearest neighbors is an algorithm that stores all available (previous) cases and uses that to predict the values based on a similarity measure. It uses 'feature similarity' to predict the values for test data/new data points. The value of the new point is assigned based on how closely it resembles other training data examples. KNN regression has two approaches. First is by calculating the average of the target of the K-nearest neighbors. Second is by computing an inverse distance weighted average of the K-nearest neighbors. KNN regression uses the same distance functions as KNN classification—Euclidean, Manhattan, and Minkowski. Initially, we try and eliminate all null values, replace missing values, so we can then use the data for the classifiers.

5 Experimental Setup

The dataset used here is Rossmann dataset, and scikit-learn library in Python was used for model selection, preprocessing, linear regression, and KNN. The matplotlib Python library was implemented for plotting graphs and data visuals. Furthermore, the sklearn.metrics module was applied for model evaluation using RMSE and MAPE.

To obtain the graph, k-fold cross-validation had been used, where $k =$ number of folds. The dataset is divided into $k = 10$ subsets, and the holdout method is repeated k times thereby improving it. By doing so, a clear and accurate prediction of sales value was obtained, then termed as the predicted value from the above graph (Fig. 5).

The histograms on the diagonals in Fig. 6 illustrate the distribution of a single variable, whereas the rest of the graphs depict the relationship between the two features. Figure 6a shows the relationship between sales and promo while Fig. 6b shows the relationship between sales and number of customers.

Even though the RMSE and MAPE with training data for both models do not show much difference, we observe a larger value for test data (Table 2). It can be observed that KNN regression is an overfitting model. Regression model score is a metric that depicts the accuracy of each of the above classifiers. It can be noticed that linear regression has a slightly better model score when compared to KNN regression. Therefore, it can be concluded that linear regression is a better model to predict sales from the given dataset.

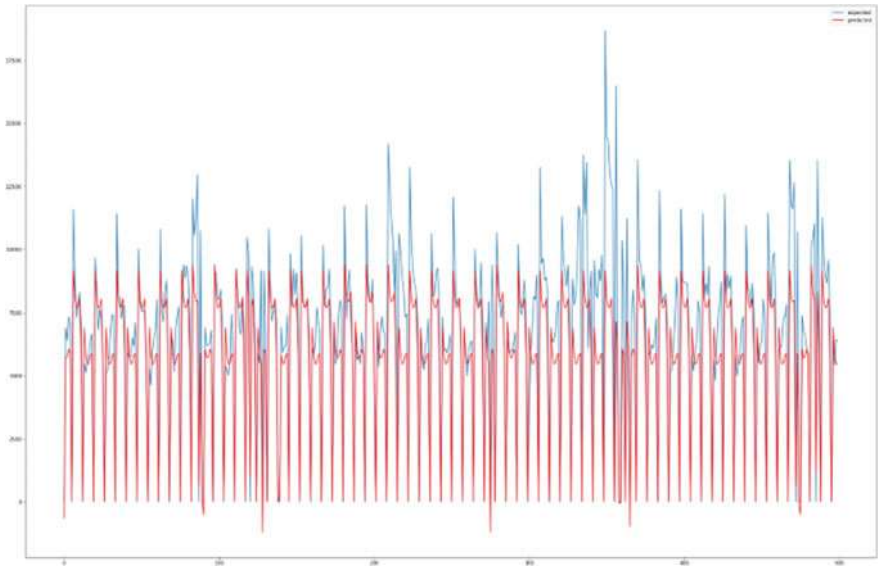

Fig. 5 Linear regression model

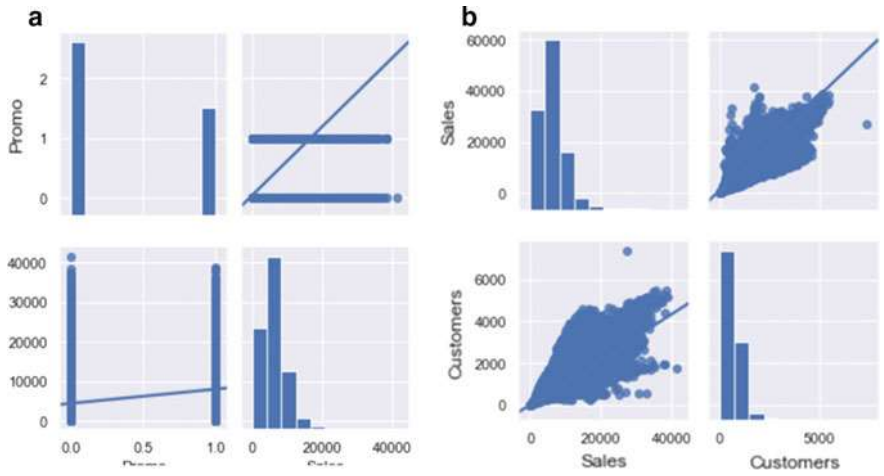

Fig. 6 (**a**) Regression line for sales and promo. (**b**) Regression line for sales and customers

Table 2 Results obtained

Model used	Model score (%)	RMSE	MAPE
Linear regression	72.19	Training data: 1742.51	Training data: 20.97
		Testing data: 1898.91	Testing data: 22.065
KNN regression	71.28	Training data: 1770.61	Training data: 21.83
		Testing data: 2546.79	Testing data: 31.40

6 Conclusion

KNN techniques are nonparametric hence mainly used in forestry problems like remote sensing. But parametric regression analysis, i.e., linear regression, has the advantage that it is easy to fit, we need to estimate only a small number of coefficients and is easy to interpret, whereas the statistical properties of KNN regression are explored lesser. In KNN regression, we observe that the difference between the training and testing error is higher than when compared to the linear regression model—which means that the KNN regression model works better for training data than testing data, thereby implying that it is an overfitting model. From this study, we can say that for the above-chosen dataset linear regression is a better model as testing and training errors for RMSE and MAPE are lesser, i.e., the difference between training and testing error is lesser for linear regression when compared to KNN regression. Therefore, it can be concluded that using linear regression we can

accurately predict sales in the future. This will help companies/organizations plan their resources better and also helps in cost optimization—maximizing profit with minimum resources.

References

1. V. Kavya, S. Arumugam, A review on predictive analytics in data mining. Int. J. Chaos Control Modell. Simul. (IJCCMS) **5**(1/2/3) (2016)
2. R.R. Shelke, R.V. Dharaskar, V.M. Thakare, Data mining for supermarket sale analysis using association rule. Int. J. Trend Sci. Res. Dev. **1**(4). ISSN: 2456-6470
3. T. Wilson, S. Asthana, Predictive Modelling for Assessing the Sales Potential of the Customer. https://www.academia.edu/28362014/Predictive_Modelling_for_Assessing_the_Sales_Potential_of_the_Customer (2016)
4. J. Gonzalez, Sales Forecasting and the Role of Predictive Analytics, (July 18, 2017), [Online]. Avaliable: https://vortini.com/blog/forecasting-predictive-analytics
5. S. Makridakis, E. Spiliotis, V. Assimakopoulos, The accuracy of machine learning (ML) forecasting methods versus statistical ones: extending the results of the M3-competition (2017)
6. M. Xue, C. Zhu, Applied research on data mining algorithm in network intrusion detection 275–277. https://doi.org/10.1109/jcai.2009.25 (2009)
7. Y.M. Khaing, M.M. Yee, E. Ei, Forecasting stock market using multiple linear regression Aung. Int. J. Trend Sci. Res. Dev. (IJTSRD) **3**(5) (2019)
8. B.M. Pavlyshenko, Machine-learning models for sales time series forecasting, Lviv, Ukraine 21–25 August 2018, pp 3–11
9. G. Nguyen, Kedia, Jai, Snyder, Ryan, Pasteur, R., Wooster, R. Sales Forecasting Using Regression and Artificial Neural Networks. (2013)
10. A. Aima, WALMART sales data analysis & sales prediction using multiple linear regression in R programming Language, [Online], Available: https://medium.com/@arneeshaima/walmart-sales-data-analysis-sales-prediction-using-multiple-linear-regression-in-r-programming-adb14afd56fb (March 19)
11. P. Mekala, B. Srinivasan, Time series data prediction on shopping mall. Int. J. Res. Comput. Appl. Robot. **2**(8), 92–97 (2014). ISSN 2320-7345
12. M. Krause-Traudes, S. Scheider, S. Rüping, Spatial data mining for retail sales forecasting, in *11th AGILE International Conference on Geographic Information Science* (2008)
13. W. Huang, Q. Zhang, W. Xu, H. Fu, M. Wang, X. Liang, A novel trigger model for sales prediction with data mining techniques. Data Sci. J. **14**, 15 (2015). https://doi.org/10.5334/dsj-2015-015
14. E. Bank, How to develop & use a regression model for sales forecasting, Updated September 26, 2017. https://bizfluent.com/how-7298496-develop-regression-model-sales-forecasting.html. Accessed 4 Oct 2019
15. Rossmann Store SalesIKaggle, Kaggle.com, 2019. [Online]. Available: https://www.kaggle.com/c/rossmann-store-sales/data. Accessed 07 Sept 2019