ABSTRACT

In today’s world where competition is cut-throat and making business decisions is increasingly difficult, the propensity to accurately make predictions is of extreme relevance. The basis of this paper is sales prediction which is a more established yet still profoundly captivating application of forecasting. Sales forecasting uses trends identified from historical data to predict future sales, enabling educated decisions including assigning or redirecting current inventory, or effectively managing future production, which forests is an ensemble learning method for classification, regression and other tasks, that functions by building a large number of decision trees at training time and producing the value that is the mean of the values of the individual trees at training time and producing the value that is the mean of the values (regression) of the individual trees, Linear regression, which is also an ensemble learning method for regression, in that it minimizes the loss function by adding regression trees using the gradient descent procedure, and Decision Trees, which essentially consists of randomizing vigorously both attribute and cut-point selection while splitting the node of a tree.

KEYWORDS: Sales Prediction, Forecasting, Ensemble Learning, Regression Trees, Decision Trees

1. INTRODUCTION

In recent years, new energy vehicles have occupied more and more market shares. If its monthly sales are predicted, it can provide reference for the production of enterprises and the formulation of national policies. For example, it would be exceptionally beneficial to be able to predict the ups and downs of a country’s economy or the fluctuations of its stock market prices. Forecasting has been done across a wide array of domains and spheres including environmental fields such as weather or even in sports performance due to the advantageous nature of prediction. The basis of this paper is sales prediction which is a more established yet still profoundly captivating application of forecasting. Sales forecasting uses trends identified from historical data to predict future sales, enabling educated decisions including assigning or redirecting current inventory, or effectively managing future production, which forests is an ensemble learning method for classification, regression and other tasks, that functions by building a large number of decision trees at training time and producing the value that is the mean of the values of the individual trees at training time and producing the value that is the mean of the values (regression) of the individual trees, Linear regression, which is also an ensemble learning method for regression, in that it minimizes the loss function by adding regression trees using the gradient descent procedure, and Decision Trees, which essentially consists of randomizing vigorously both attribute and cut-point selection while splitting the node of a tree. The study aims to compare three machine learning methods for sales prediction in the food industry: Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Radial Basis Function Network (RBFN). The methods were compared in
terms of their prediction accuracy on daily sales in a food store department. The performance of the models was determined using the performance measures: Mean Average Percentage Error (MAPE) and Root Mean Squared Error (RMSE). Low accuracy Little amount data is used. We apply three algorithms and predict Walmart dataset. Historical sales data for 45 Walmart stores located in different regions. Each store contains many departments, and participants must project the sales for each department in each store. To add to the challenge, selected holiday markdown events are included in the dataset.

2. RELATED WORKS

Popular approaches to the prediction problem are machine learning approaches using Artificial Neural Networks (ANNs) and statistical approaches using Autoregressive Integrated Moving Average (ARIMA). Both methods when applied to stock predictions and sales forecasting have had moderate success. ANNs have typically displayed good performance in forecasting due to their capacity to characterize non-linear data with very good accuracy. However, their results could be further improved through better selection of hyperparameters. A combination of Seasonal Trend Decomposition using Loess and Autoregressive Integrated Moving Average (STL+ ARIMA) has also been used in time series forecasting and generated good results. The work of Pao et al, proved to be a significant steppingstone to the use of machine learning algorithms on which this paper is based. Another solid approach to sales prediction employed an extreme learning machine that forecasted sales in the fashion retail industry. The results of this work, published by Sun et al, contributed to the hyperparameter selection proposed by this paper. The paper entails three algorithms namely, Random Forest, Linear Regression, and Decision Trees, that are executed on the Walmart dataset. The algorithms were implemented using Python running on Jupyter Notebooks in the Anaconda distribution. The performance of each algorithm was compared to highlight the best results.

3. PROPOSED WORK

a. DATASET

The dataset comes from the Kaggle platform and consists of data from an American retail organization, Walmart Inc. The dataset was used for a machine learning competition in 2014 [10]. It comprises data from 45 Walmart department stores mainly centered around their sales on a weekly basis. The dataset has 282,452 entries that will be used for training the models. Each entry has attributes as follows: the associated store (recorded as a number), the corresponding department (81 departments, each entered as a number), the date of the starting day in that week, departmental weekly sales, the store size, and a Boolean value specifying if there is a major holiday in the week. The major holidays being one of Thanksgiving, Labor Day, Christmas or Easter. Along with the aforementioned attributes is a parallel set of features for each entry including Consumer Price Index, unemployment rate, temperature, fuel price, and promotional markdowns. Since there is no test-set provided, they are generated from the given training data for cross-validation, and final testing.

b. ML Machine Methods

Three forecasting models were constructed in this research on the following algorithms: Random Forest, Gradient Boosting, and Extremely Randomized Trees (Extra Trees). Other algorithms such as Naïve Bayes and Adaptive Boosting were scrutinized, but their performances were not up to the mark and insights were trivial, so they will not be considered herein. All models were implemented in Python 3.7. on the Anaconda distribution using Jupyter Notebooks. But we applied Random forest, decision tree and Linear regression.
c. Random Forest:

The Random Forest architecture is best described by. As more trees are grown, the Random Forest algorithm adds more randomness to the model. It searches for the best feature amidst a random subset of features in place of searching for the most relevant feature while splitting a node. This results in more accurate model as it leads to a much greater diversity. Thus, in Random Forest, only a random subset of the features is considered by the algorithm for diverging a node. Trees can be made more random by using random thresholds for each feature instead of searching for the best thresholds (like a normal decision tree does). The features used for training the model were week number, store number, department number, the holiday flag, Consumer Price Index, unemployment rate, temperature, fuel price and store size. The algorithm was carried out using Python’s Random Forest Regressor function present in the scikit-learn class. In the Python implementation, Mean Absolute Error (MAE), mean-squared error (MSE) and R2 score are calculated for the predicted values.

Fig. 1.1. Random Forest

![Random Forest Diagram]

The Extra Trees and Random Forest algorithms are almost the same. In the Random Forest algorithm, the tree splitting phenomenon is deterministic in nature whereas in the case of Extremely Randomized Trees, the split of the trees is completely random. In other words, during the process of splitting, the algorithm chooses the best split among random splits in the selected variable for the current decision tree. The features employed are like the ones used in the previous algorithms. Python’s Extra Trees Regressor function from the scikit-learn class was used to execute the algorithm, and the various performance metrics calculated for the previous methods are evaluated and reported. Figure 5 shows the comparison of the predicted values and the actual values of the weekly sales with the hyperparameters set at the optimized values.

d. Decision Trees

![Comparison of Predicted and Actual Weekly Sales for Extra Trees]

Fig. 1.2. Comparison of Predicted and Actual Fig

![Weekly Sales for Extra Tress]

Fig.1.2. Weekly Sales for Extra Tress

e. Linear Regression

Linear regression is a common Statistical Data Analysis technique. It is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line. Linear regression quantifies the
relationship between one or more predictor variables and one outcome variable. For example, linear regression can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). In simple linear regression a single independent variable is used to predict the value of a dependent variable. In multiple linear regression two or more independent variables are used to predict the value of a dependent variable. The difference between the two is the number of independent variables. In both cases there is only a single dependent variable.

4. Arima

ARIMA stands for Auto Regressive Integrated Moving Average. This is one of the most popular models used for Time Series Forecasting. It makes use of 3 variables: P: stands for Periods to lag. It denotes the number of earlier periods of our time series to use for forecasting. This is the Auto Regression part of the calculation. D: stands for Differencing. In an ARIMA model, we transform a time series into stationary one (without any trend or seasonality). It refers to the number of differencing transformations required by the time series to get stationary. This is the Integrated portion of the calculation. Q: stands for Error component to lag. It is a part of the time series not explained by trend or seasonality. This is the Moving Average part of the calculation.

5. Architecture

![Figure 1.5. System Architecture](image1)

![Figure 1.6. Exploratory Data Analysis](image2)

**FIG 1.7. RANDOM FOREST REGRESSOR**
6. CONCLUSION

This paper dealt with the implementation of three algorithms namely, Random Forest, Linear Regression, and Extra Trees, on the Walmart dataset and a comparative analysis was carried out to determine the best algorithm. Random Trees was confirmed to be a very effective model in forecasting sales data. Extra Trees, an extension of Random Forest, also showed very good accuracy for the best implementations. These algorithms could possibly produce even better results if they are provided with better hardware electronics like Graphics Processing Units (GPUs). Future work would include the Extra Trees model being developed to consider sparse promotional markdown data and moving holidays. It would also involve the fine-tuning of the hyperparameters of the models to improve the accuracy of prediction. Future work could also entail combining the models to produce an ensemble training model that could represent even the tiniest details present in the data. With the development of deep learning techniques, the results of this research could be further improved soon using more complex and multilayer ANNs. This work shows that there are highly efficient algorithms to forecast sales in big, medium or small organizations, and their use would be beneficial in providing valuable insight, thus leading to better decision-making.

REFERENCE


