Drug Store Sales Prediction using Machine Learning

Raaghav Sarraf
Birla Institute of Technology
Mesra,Ranchi,India.
raghavsaraf15@gmail.com

Dewang Sultania
Birla Institute of Technology
Mesra,Ranchi,India.
dewangsultania@gmail.com

***Abstract*—We applied Machine Learning techniques to real-world problem of predicting store sales. This kind of prediction enables store managers to create effective staff schedules that increase productivity and motivation. We used feature selection, feature processing and model selection to improve our result. In view of the nature of our problem, Root Mean Squared Percentage Error (RMSPE) is used to measure the prediction accuracy.**

Keywords—Machine Learning; Sales; Linear Regression; CART; K-Nearest Neighbors; Random Forest.

I. Introduction

Nowadays medical-related sales prediction is of great interest; with reliable sales prediction, medical companies could allocate their resources more wisely and make better profits. Rossmann is a chain drug store that operates in 7 European countries. We obtained Rossmann 1115 Germany Stores sales data from Kaggle.com.The goal of this project is to have reliable sales prediction for each store for up to six weeks in advance. The input to our learning algorithm contains many impacting factors like sales, store type, date, promotion, competition, etc. Algorithms like Linear Regression, K-Nearest Neighbors, Classification and Regression Trees and Random Forest Regressor were used to predict sales.

II. Methodology

A. Dataset and Features

The training data provided contained the following features:

- Date
- Day of week
- Store ID
- Customers
- Open
- State Holiday
- School Holiday
- Store Type
- Assortment
- Competition Distance
- Competition Open Since Month/Year
- Promo
- Promo2
- Promo2 Since Year/Week
- Promo Interval

The Response Variable we needed to predict was SALES

B. Pre-processing and Feature Extraction Methods

Features like Store, Competition Distance, Promo, Promo2 and School Holiday were used directly. For the features StoreType, Assortment and State Holiday which contained the values ‘a’, ‘b’, ‘c’ and ‘d’ a mapping was used which mapped ‘a’ to 1, ‘b’ to 2, ‘c’ to 3 and ‘d’ to 4.Now features like Month, Day and Year and week of Year was calculated from the given Date feature. Now feature Competition Open was calculated with the help of features Competition Open Since Year/Month, Year and Month using the formula:

$$\text{CompetitionOpen}=12*(\text{Year}-\text{CompetitionOpenSinceYear})+(\text{Month}-\text{CompetitionOpenSinceMonth})$$

A new feature PromoOpen was introduced which specified the time duration for which the Promo was active. We also classified Months of the year as a promo or non-promo month by adding a new feature IsPromoMonth.

Now different Machine Learning Models like Linear Regression, K-Nearest Neighbors, Classification and Regression Trees and Random Forests were trained using 80% of the data and the rest was used for validation.

III. Related work

Linear Regression : Linear Regression is a simple model that combines a specific set of input values(x) the solution to which is the predicted output for that set of input values y. Both input and output are numeric.

$$y= a_0+\Sigma\ (a_i x_i)$$

K-Nearest Neighbors: KNN makes predictions using the training set directly. Predictions are made for new data points by searching through the entire dataset for k most similar instances and summarizing the output variable for those k instances. For regression this is the mean output variable.

Classification and Regression Trees: Classification and Regression Trees or CART for short is a term introduced by Leo Breiman to refer to Decision Tree algorithms that can be used for classification or regression predictive modeling problems. The representation for the CART model is a binary tree. Each node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric). The leaf nodes of the tree contain an output variable (y) which is used to make a prediction.

RandomForestRegressor: It undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps required for data exploration. It is a type of ensemble learning method where a group of weak models combine to form a powerful model. They operate by constructing multiple decision trees at the training time and outputting the mean of the individual trees.

Residual Sum of Squares= $\Sigma\ (y_i\text{-}y_L)^2 + \Sigma\ (y_i\text{-}y_R)^2$

Where y_L = mean y value of the left node

y_R= mean y value of the right node

IV. Results

The performance metric used was Root Mean Square Percentage Error (RMSPE) which is calculated using the following formula:

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

The results obtained by using the different algorithm are as mentioned in Table I.

TABLE I. Results

S.No.	Algorithms and RMSPE Values		
	Name of Algorithm	*RMSPE*	*Number of Estimators*
1.	Linear Regression	0.53041	NA
2.	K-Nearest Neighbor	0.46151	5
3.	CART	0.32582	NA
4.	Random Forest Regression	0.13595	10
5.	Random Forest Regression	0.13080	30
6.	Random Forest Regression	0.12875	50
7.	Random Forest Regression	0.12817	100

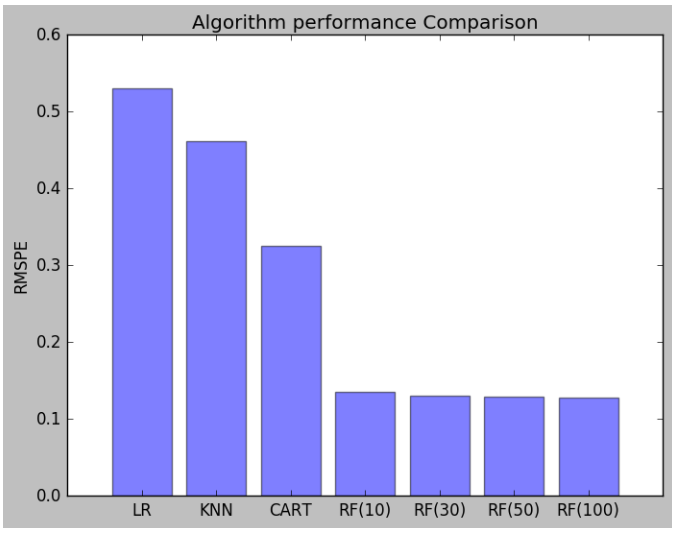

Fig. 1. Comparison of Algorithm.

References

[1] https://www.kaggle.com/c/rossmann-store-sales

[2] Tom M. Mitchell, “Machine Learning,” McGraw Hill Education(India) Private Limited

[3] A.D’yakonov, “Supermarkes clients behavior forecasting by weighted methods of probability and density estimations,” Business InformaticsNo1(27)-2014,MSU ,2014

[4] Andrew Y. Ng, “Advice for applying Machine Learning, Stanford University.” http://cs229.stanford.edu/materials/ML-advice.pdf

[5] Breiman, L. (2001). “Random Forests,” Machine Learning, 45(1), 5-32.

[6] Chen, Tianqi. “Introduction to Boosted Trees.” (2014):n. Pag. Https://homes.cs.washington.edu/tqchen/pdf/BoostedTree.pdf. University of Washington