Sales Forecasting for Retail Chains

Ankur Jain[1], Manghat Nitish Menon[2], Saurabh Chandra[3]
A53097130[1], A53097652[2], A53104614[3]
{anj022[1], mnmenon[2], sbipinch[3]}@eng.ucsd.edu

***Abstract*—This paper presents a use case of data mining for sales forecasting in retail demand and sales prediction. In particular, the Extreme Gradient Boosting algorithm is used to design a prediction model to accurately estimate probable sales for retail outlets of a major European Pharmacy retailing company. The forecast of potential sales is based on a mixture of temporal and economical features including prior sales data, store promotions, retail competitors, school and state holidays, location and accessibility of the store as well as the time of year. The model building process was guided by common sense reasoning and by analytic knowledge discovered during data analysis and definitive conclusions were drawn. The performances of the XGBoost predictor were compared with those of more traditional regression algorithms like Linear Regression and Random Forest Regression. Findings not only reveal that the XGBoost algorithm outperforms the traditional modeling approaches with regard to prediction accuracy, but it also uncovers new knowledge that is hidden in data which help in building a more robust feature set and strengthen the sales prediction model.**

Keywords: Sales Prediction, Random Forest Regression, Linear Regression, XGBoost, Time Series, Gradient Boosting.

I. Introduction

Retail is one of the most important business domains for data science and data mining applications because of its prolific data and numerous optimization problems such as optimal prices, discounts, recommendations, and stock levels that can be solved using data analysis methods. The usual problems tackled by data mining applications are Response Modeling, Recommendations systems, Demand prediction, Price Discrimination[12], Sales Event Planning, and Category Management[1]. Accurate forecasting of customer demand remains a challenge in today's competitive and dynamic business environment and minor improvements in predicting this demand helps diversified retailers lower operating costs while improving sales and customer satisfaction

Predicting the right demand at each retail outlet is crucial for the success of every retailing company because it helps towards inventory management, results in better distribution of produce across stores, minimizes over and under stocking at each store thereby minimizing losses, and most importantly maximizes sales and customer satisfaction[3]. Due to the high stakes involved with demand prediction, it becomes a vital problem to solve for every retail company[19]. Further, demand can depend on a variety of external factors like competition, weather, seasonal trends, etc and internal actions like promotions, sales events, pricing, assortment planning etc, adding to the complexity of the problem. Consequently, the modeling of demand prediction taking into account all of the factors per retail outlet becomes essential for every retail company. Therefore, this paper proposes an approach for demand and sales prediction for retail at each outlet.

Having fixed on the data mining problem of Sales prediction at each outlet of a retailing company, Rossmann - Germany's second-largest drug store chain with 3,000 stores in Europe, was chosen as the retailing company for data to develop the predictive model on. The data was available through their first Kaggle competition. In the competition Rossmann challenged to predict 6 weeks of daily sales for 1115 of their stores located across Germany. This is a crucial problem for Rossmann as currently their store managers are tasked with predicting their daily sales for up to six weeks in advance. Since store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality and with thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results has been quite varied. Therefore reliable sales forecasts will enable their store managers to create effective staff schedules that increase productivity and motivation.

II. Related Work

1) **Microsoft Time Series Algorithm**[2]**:** The Microsoft Time Series algorithm provides regression algorithms that are optimized for the forecasting of continuous values over time. The major advantage of the Time Series Algorithm is that it does not require additional columns of new information as input to predict a trend whereas other algorithms based on decision trees them. A time series model can predict trends based only on the original dataset that is used to create the model. Any new data added to the model when making a prediction is automatically incorporated into the trend analysis. Another unique feature of the Microsoft Time Series algorithm is that it can perform cross prediction. The algorithm can be trained with two separate, but related, series, and the resulting model created can predict the outcome of one series based on the behavior of the other series. For example, the observed sales of one product can influence the forecasted sales of another product. How it Works: The Microsoft Time Series algorithm uses both methods, ARTXP (Autoregressive Tree Models with Cross Prediction) and ARIMA (Autoregressive Integrated Moving Average), and blends the results to improve prediction accuracy. The ARTXP algorithm can be described as an autoregressive tree model for representing periodic time series data. The ARIMA algorithm improves long-term prediction capabilities of the Time Series algorithm.

2) **Spatial data mining for retail sales forecasting[13]:** This paper presents a use case of spatial data mining for aggregate sales forecasting in retail location planning. Support Vector Regression (SVR) is the technique used to design a regression model to predict probable turnovers for potential outlet-sites of a big European food retailing company. The forecast of potential sites is based on sales data on shop level for existing stores and a broad variety of spatially aggregated geographical, socio-demographical and economical features describing the trading area and competitor characteristics. The model was built from a-priori expert knowledge and by analytic knowledge which was discovered during the data mining process. To assess the performance of this SVR-model, it was compared to the traditional state-of-the-art gravitational Huff-model. The spatial data mining model was found to outperform the traditional modeling approach with regard to prediction accuracy. Support Vector algorithms are specially designed to minimize the expected classification error by minimizing both the empirical error and complexity. SVR works on almost the same principles as the Support Vector Classification. The SV-approach searches for the linear classifier which separates the positive from the negative instances of the training set by maximizing the margins. The margin is the distance between the separating line and the nearest data points (the Support Vectors). This linear classifier is called the Optimal Separating Hyperplane. Kernel functions are used to handle instances where the data points are not linearly separable which works by transforming the input space containing the training instances into a new, higher-dimensional feature space, in which it becomes possible to separate the data.
3) **A Novel Trigger Model for Sales Prediction with Data Mining Techniques[8]:** This paper describes an approach on how to forecast sales with higher effectiveness and more accurate precision. The data used in this approach focuses on online shopping data in the Chinese B2C market. The paper delves into e-commerce[17] and applies real sales data to several classical prediction models, aiming to discover a trigger model that could select the appropriate forecasting model to predict sales of a given product. The paper aims to effectively support an enterprise in making sales decisions in actual operations. The approach involves manipulating raw data into available forms and then a trigger model is proposed to do the classification. The classification result indicate the best prediction model for each item. Finally, by use of the most appropriate model, the prediction is accomplished. The features used are - CV Sales (CVS): coefficient of variation for sales , CV Attention (CVA): coefficient of variation for attention , Sold Price Variation (SPV): the variation of sold price. This approach involves applying two typical forecasting models and several dimensions to the trigger model through training and testing the classification model with real sales data and focuses on the correlation of two subjects and ignores the causal relationship between them.

III. Dataset Exploration

A. Data Description

1) Store - Each store in the dataset has a unique ID associated with it
2) Sales - The turnover for a store on a given day
3) Customers - the number of customers who visited the store on a given day
4) Open - indicates whether the store was open(0) or closed(1)
5) StateHoliday - indicates a state holiday. There are 4 classes -> a = public holiday, b = Easter holiday, c = Christmas, 0 = None
6) SchoolHoliday - indicates if the was affected by the closure of public schools. These holidays vary from state to state.
7) StoreType - differentiates between 4 different store models: a, b, c, d. Different kinds of stores sell different products
8) Assortment - describes an assortment level: a = basic, b = extra, c = extended. Indicates the variety in items the store sells.
9) CompetitionDistance - distance in meters to the nearest competitor store
10) CompetitionOpenSince[Month/Year] - The year and month the nearest competitor was opened
11) Promo - indicates whether a store is running a promo on that day : 1 indicates a promo, 0 indicates no promo
12) Promo2 - a continuing and consecutive promotion for some stores: 1 = store is conducting the promo, 0 = store is not conducting the promo
13) Promo2Since[Year/Week] - the year and week when the store started participating in Promo2
14) PromoInterval - Promo2 runs during certain months of the year, this field indicates this event.
15) DayOfWeek - Varies from Monday to Sunday. Most stores are closed on Saturday and Sunday.

The Rossmann Data contains information about 1115 stores from 1st Jan 2013 to 31st July 2015 (942 days). In total we have 1017209 entries[16].

Fig 1 depicts the histogram of the mean sales per store when the stores weren't closed. From this graph we can infer that most stores have very similar sales and that there is a small percentage of outliers.

Fig 2 shows how the number of customers and sales average vary based on the day of the week taken into consideration. As expected average sales tend to be more on the Sunday as compared to other days. However, the magnitude of customers visiting a store on weekends tends to be less due to the fact that most stores are closed on Sunday.

B. Detecting Trends:

Using services such as Google Trends makes it easy to visualize the data and draw conclusions from it. It helps by giving information on external conditions which explain certain outliers in the training data. For example analyzing Weather data[14] gives a good idea (heavy snowfall) on why certain stores in a geographical area had lower sales at certain

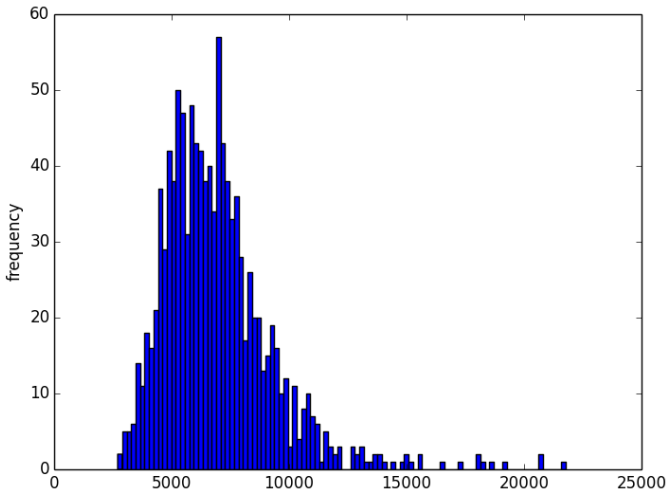

Figure 1: Mean Sales when store was not closed

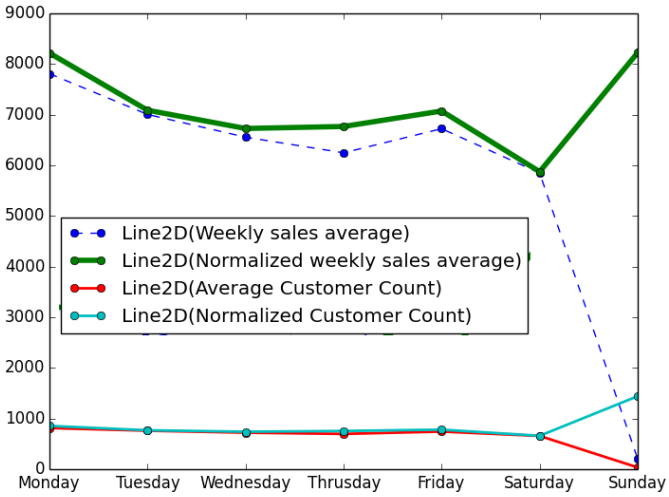

Figure 2: Mean Sales/Customers by day of week

weeks of the year. It also indicated that Pre-holiday sales are higher in stores as opposed to sales on a normal day. This could be due to promo offers being run during the holiday period. Also state holidays vary by date from state to state

In addition data was looked into for the following Geographical and Social Factors -

1) Accessibility
2) Store Location
3) Store Competition
4) Population Density

C. Missing Data:

There are 180 stores missing 184 days of data in the middle of the series between 1 July 2014 to 31 Dec 2014

D. Feature Relative Importance Estimation

Decision tree is known for its ability to select "important" features among many and ignore (often irrelevant) others. In addition, decision tree gives an explicit model describing the relationship between features and predictions, thus easing model interpretation. Random forest, as an ensemble of trees, inherits the ability to select "important" features. However, it does not produce an explicit model. Instead, the relationship between features and activity of interest is hidden inside a "black box". Nonetheless, a measure of how each feature contributes to the prediction performance of random forest can be calculated in the course of training. When a feature that contributes to prediction performance is "noised up" (e.g., replaced with random noise), the performance of prediction is noticeably degrade. On the other hand, if a feature is irrelevant, "noising" it up should have little effect on the performance. Thus, we can estimate the relative importance of features according to the following procedure[18]. As each tree is grown, it makes predictions on the OOB (out-of-bag) data for that tree. At the same time, each feature in the OOB data is randomly permuted, one at a time, and each such modified dataset is also predicted by a tree. At the end of the model training process, RMPSEs are calculated based on the OOB prediction as well as the OOB predictions with each features permuted. Let M be the RMPSE based on the OOB prediction and M_j the RMSE based on the OOB prediction with the j^{th} feature permuted. Then the measure of importance for the j^{th} feature is simply $M_j - M$.

Table I: Relative Importance

Variable	Relative Importance	Scaled Importance
Store	11101946.000000	1.000000
Promo	1639790.750000	0.147703
DayOfWeek	713757.875000	0.064291
Month	375840.531250	0.033854
CompetitionSinceMonth	270933.468750	0.024404
CompetitionSinceYear	237843.515625	0.021424
StoreType	195947.984375	0.017650
CompetitionStrength	181692.234375	0.016366
Promo2 SinceWeek	151932.953125	0.013685
Assortment	121404.148438	0.010935
Year	117898.539062	0.010620
Promo2SinceYear	114260.554688	0.010292
SchoolHoliday	62071.664062	0.005591
Promo2	34230.789062	0.003083
SundayStore	27033.765625	0.002435
StateHoliday	12047.477539	0.001085

E. Features

A plethora of features were considered during experimentation but finally the features listed below had the greatest impact on the model

1) Mean Sale per store per day of the week for a particular storeType
2) Day of the week
3) Store Type
4) Assortment
5) log(Distance from nearest competition)
6) Promo
7) Month
8) Year
9) School Holiday

Fig 3 showcases the comparison of sales for a store (262) on Sunday vs the other days of the week. The graph justifies the

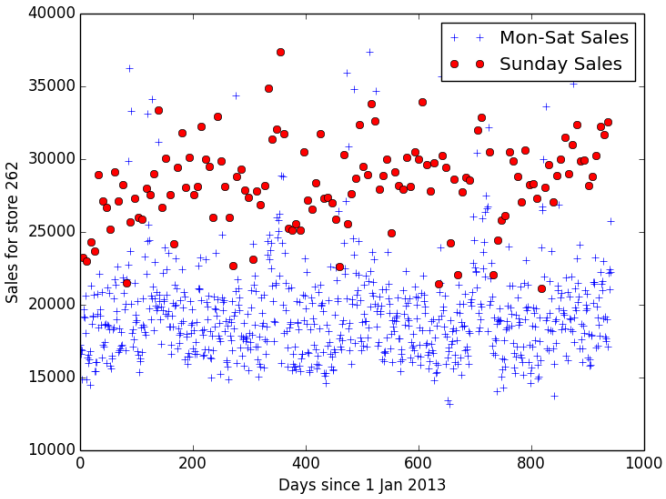

Figure 3: Sales of Store 262

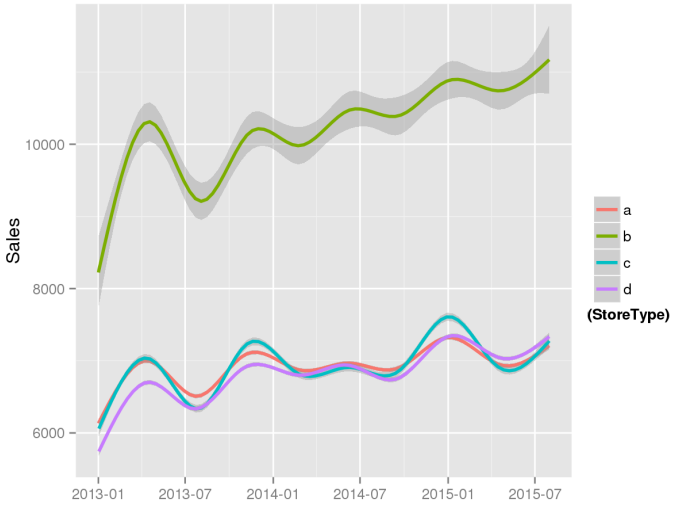

Figure 4: Sales filtered by StoreType

assumption that sales on Sunday are better than the rest of the week.

Fig 4 and Fig 5 show that there is a direct correlation between the number of customers and the store sales. From the distribution in the graph, it is very evident that the sales are much higher if the store is of type ‘b’ as compared to the other store types ‘a’, ‘c’, and ‘d’. In addition the number of customers per store over a time period is highest for store b justifying our assumption

Fig 6 and Fig 7 indicate that the assortment of products being sold at a store logically impacts sales as customers tend to buy items according to their needs and availability. From the distribution, it is very evident that the sales are highest if the store has an assortment level of ’a’ which implies ’extra’, followed by the assortment level ’c’ which indicates is ‘extended’. The sales are least if the assortment level is ‘basic’. This inference matches with the common sense reasoning that sales increases as the assortment of products increases and customers prefer visiting stores with greater assortment.

Identifying how sales are linked to seasonal trends goes a

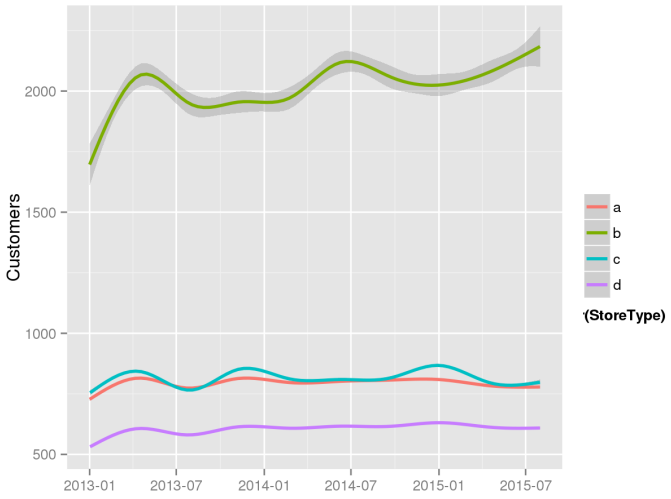

Figure 5: Number of Customers filtered by StoreType

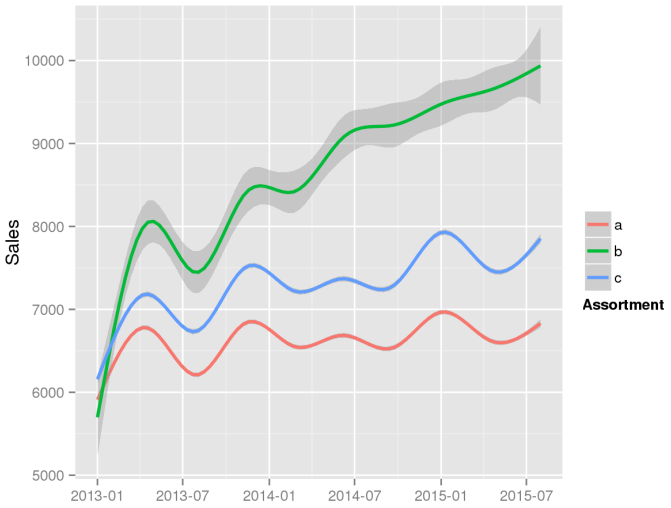

Figure 6: Sales filtered by Assortment

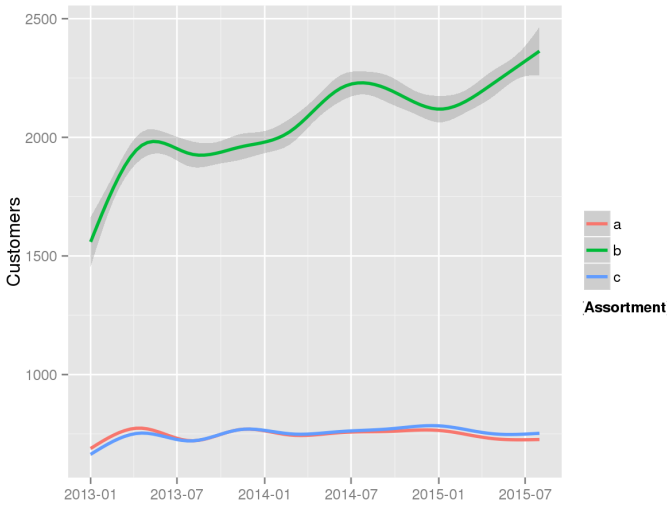

Figure 7: Number of Customers filtered by Assortment

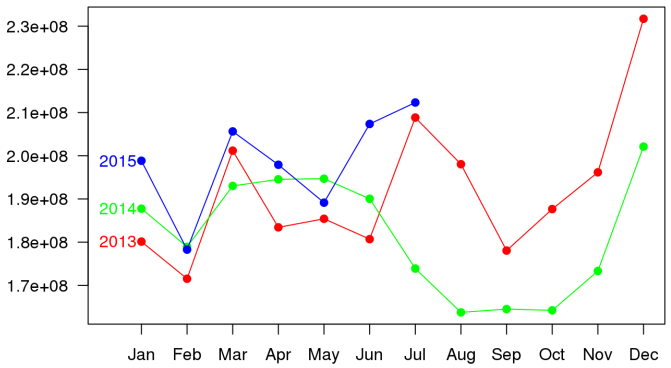

Figure 8: Seasonal Sales

long way towards helping retailers generate higher revenue by ensuring they are well stocked at the right time . Fig 8 shows that sales are highest during Christmas and around New Year and experience a significant dip right after. Sales during the months of July and August are quite high. The graph shows that seasonal sales trends have not changed significantly over a three year period.

IV. Experiments

A. XGBoost: Extreme Gradient Boosting

While looking at better techniques for data analysis and forecasting online, we came across XGBoost which gives much better performance results than Linear Regression or Random Forest Regression. XGBoost[5] or Extreme Gradient Boosting is a library that is designed, and optimized for boosted (tree) algorithms. The library aims to provide a scalable, portable and accurate framework for large scale tree boosting. It is an improvement on the existing Gradient Boosting technique.

Gradient Boosting:

Gradient boosting[7] is a machine learning technique for regression and classification problems, which produces a prediction model in the form of weak prediction models, typically decision trees. Boosting can be interpreted as an optimization algorithm on a suitable cost function. Like other boosting methods, gradient boosting combines weak learners into a single strong learner, in an iterative fashion.

XGBoost is used for supervised learning problems, where we use the training data x to predict a target variable y. The regularization term controls the complexity of the model, which helps us to avoid overfitting. XGBoost is built on a Tree Ensemble model which is a set of classification and regression trees (CART). We classify the members of a family into different leaves, and assign them the score on corresponding leaf. The main difference between CART and decision trees is that in CART, a real score is associated with each of the leaves in addition to the decision value.This gives us richer interpretations that go beyond classification. It consists of 3 steps -

1) Additive Training In this stage we define the parameters of trees which are those functions that contain information about the structure of the tree and the leaf score. This is optimized by using an additive strategy: fix what we have learned, add a new tree at a time.
2) Model Complexity The complexity of the tree acts as our regularization parameter and helps decide how to penalize certain cases.
3) Structure Score This score is gives information on the best split conditions while taking the model complexity into account. The first split of a tree will have more impact on the purity and the following splits focus on smaller parts of the dataset which have been misclassified by the first tree.

B. Linear Regression

The simple linear regression[6] model is

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

where intercept $\beta_0, and\, slope\, \beta_1$ are unknown constants and ϵ is a random error component. The errors are assumed to have mean zero error and unknown variance σ^2 Additionally we assume that the errors are un-correlated. Regressor x is controlled by the data analyst and measured with negligible error while the response y is a random variable. The mean of this distribution is

$$E(y|x) = \beta_0 + \beta_1 x \tag{2}$$

and variance is

$$Var(y|x) = Var(\beta_0 + \beta_1 x + \epsilon) = \sigma^2 \tag{3}$$

This the mean of y is a linear function of x although the variance of y does not depend on the value of x. Parameters $\beta_0, and\, \beta_1$ are known as regression coefficients.

C. Random Forest Regression

Random forest, which was first proposed by Breiman[4] , is an ensemble of B trees $T_1(X), ..., T_B(X)$, where $X = x_1, ..., x_p$ is a p-dimension vector of features. The ensemble produces B outputs $\hat{Y_1} = T_1(X), ..., \hat{Y_B} = T_B(X)$ where $\hat{Y_b}, b = 1, ..., B$, is the prediction value for a sequence by the b^{th} tree. Outputs of all trees are aggregated to produce one final prediction, $\hat{Y}$. For regression problems, $\hat{Y}$ is the average value of the individual tree predictions. Given data on a set of n sequences for training, $D = (X_1, Y_1), ..., (X_n, Y_n)$, where X_i, i = 1, ..., n, is a vector of features and Y_i is experimental validated efficacy value, the training procedure are as follows[18]:

1) From the training data of n sequences, draw a bootstrap sample (i.e., randomly sample, with replacement, n sequences).
2) For each bootstrap sample, grow a tree with the following modification: at each node, choose the best split among a randomly selected subset of mtry (rather than all) features. Here mtry is essentially the only tuning parameter in the algorithm. The tree is grown to the maximum size (i.e., until no further splits are possible) and not pruned back.
3) Repeat the above steps until (a sufficiently large number) B such trees are grown

V. Results

Test set contains entries from 1115 stores from 1st Aug 2015 to 17th Sep 2015 (48 days). In total there are 41088 entries for all stores. We evaluated our model based on RMPSE:

$$RMPSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \quad (4)$$

where y_i denotes the sales of a single store on a single day and $\hat{y}_i$ denotes the corresponding prediction. Any day and store with 0 sales is ignored in scoring.

Table II: Results

Model	RMPSE on Test Set
Mean of each DayOfWeek	0.18968
Linear Regression	0.15672
Random Forest Regression	0.13198
XGBoost	0.10532

VI. Conclusion

In this project we have performed sales forecasting for stores using different data mining techniques. The task involved predicting the sales on any given day at any store. In order to familiarize ourselves with the task we have studied previous work in the domain including Time Series Algorithm as well as a Spatial approach. A lot of analysis was performed on the data to identify patterns and outliers which would boost or impede the prediction algorithm. The features used ranged from store information to customer information as well as socio-geographical information. Data Mining methods like Linear Regression, Random Forest Regression and XGBoost were implemented and the results compared. XGBoost which is an improved gradient boosting algorithm was observed to perform the best at prediction. With efficiency being the way forward in most industries today, we aim to expand our solution to help stores improve productivity and increase revenue by taking advantage of Data Analysis.

VII. Future Work

Sales prediction plays a vital role in increasing the efficiency with which stores can operate as it provides details on the traffic a store can expect to receive on a given day. In addition to just predicting the expected sales, there are other data which can be mined to highlight important trends and also improve planning. Briefly they are -

1) Advertisement : Identifying which customers will react positively to certain ad's[15] and offers to ensure they receive them. Conversely identifying customers who do not like certain offers will help reduce sending out unnecessary offers.
2) Recommendations: Once the category of products a customer is interested in is identified, he can be recommended other products[9] he may like thereby increasing sales.
3) Predicting Demand : In addition to predicting sales, predicting the demand[11] is another solution which would immensely benefit stores. Prior knowledge of what products will be in demand will help stores stock up on the right items.
4) Customer Based Pricing : This solution involves identifying the appropriate discounts for different items so as to maximize revenue[10]. Identifying the right product will help generate profit as well as clear excess stock.
5) Holiday / Extended Sale Planning : Involves identifying the best products to offer discounts or promos on during holidays by predicting demand. In addition finding out the best time period to offer discounts will benefit stores as well as the customers.
6) Product Classification : Classifying products into a single category will help stores offer the best products to customers. Stores can avoid stocking up on redundant products as well as those that customers may not buy together.

References

[1] Shirley Coleman Ahlemeyer Stubbe, Andrea. A practical guide to data mining for business and industry. *John Wiley and Sons*, 2014.

[2] P. Mekala B. Srinivasan. Time series data prediction on shopping mall. In *International Journal of Research in Computer Application and Robotics*, Aug 2014.

[3] Gordon S. Linoff Berry, Michael JA. Data mining techniques: for marketing, sales, and customer relationship management. *John Wiley and Sons*, 2004.

[4] L. Breiman. Random forest. *Mach. Learn.*, 4:5–32, 2001.

[5] Tianqi Chen. Xgboost https://github.com/tqchen/xgboost.

[6] Douglas C Montgomery et all. *Introduction to Linear Regression Analysis*. Number 5. Wiley, 2012.

[7] Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, pages 367–378, 2002.

[8] et al. Huang, Wenjie. A novel trigger model for sales prediction with data mining techniques. *Data Science Journal*, 14, 2015.

[9] et al. Jannach, Dietmar. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[10] Romana J. Khan and Dipak C. Jain. An empirical analysis of price discrimination mechanisms and retailer profitability. *Journal of Marketing Research*, 42.4:516–524, 2005.

[11] A. Gürhan Kök and Marshall L. Fisher. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, 55.6:1001–1021, 2007.

[12] M Nakanisihi LG Cooper. *Market Share Analysis*, 2010.

[13] Simon Scheider et. all Maike Krause-Traudes. Spatial data mining for retail sales forecasting. *11th AGILE International Conference on Geographic Information Science*, 2008.

[14] Andrew G Parsons. The association between daily weather and daily shopping patterns. *Australasian Marketing Journal (AMJ)*, 9.2:78–84, 2001.

[15] Nicholas J. Radclifte and Rob Simpson. Identifying who can be saved and who will be driven away by retention activity. *Journal of Telecommunications Management*, 1.2, 2008.

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

[17] et al Schroeder, Glenn George. System for predicting sales lift and profit of a product based on historical sales information. *U.S. Patent No. 7,689,456*, 2010.

[18] A. Liaw et. all V. Svetnik. Random forest: a classification and regression tool for compound classification and qsar modeling. *J. Chem. Inf. Comput. Sci.*, 43, 2003.

[19] Wayne L. Winston. *Analytics for an Online Retailer: Demand Forecasting and Price Optimization*. Wiley, 2014.