Rossmann Store Sales Prediction

Machine Learning Final Project

Neela Avudaiappan

School of Computing

Clemson University

Clemson, USA

navudai@g.clemson.edu

Jeyes Andrew Martin

School of Computing

Clemson University

Clemson, USA

jeyesm@gmail.com

*Abstract*— Rossmann is a drug store operating all over Europe, a kaggle competition opened up which required the participants to predict the sales of 1115 stores for the next 12 months. We were given the sales data of the last 12 months of these stores as well as the store information with respect to the promotions that are run on the store as well as the competitor around the store. The sales prediction would help the manager’s better plan the store.

Keywords— stores; promotions

# Introduction

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

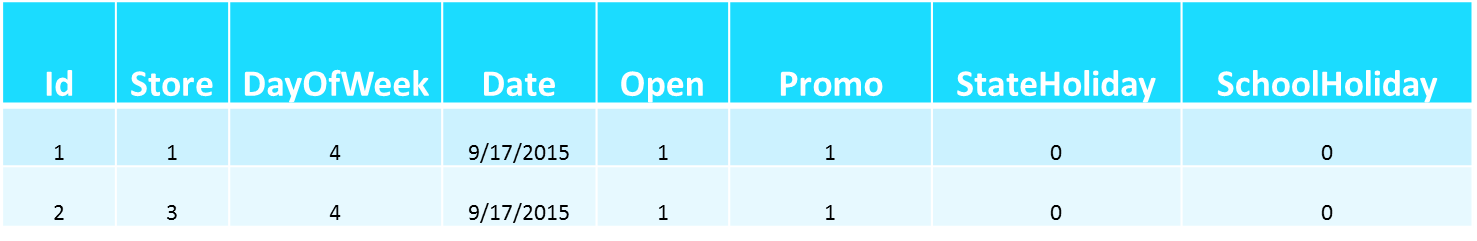
In their first Kaggle competition, Rossmann is challenging you to predict 6 weeks of daily sales for 1,115 stores located across Germany. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation. By helping Rossmann create a robust prediction model, you will help store managers stay focused on what’s most important to them: their customers and their teams!

This e.

# Available Data

The following are the available data:

* **Id** - an Id that represents a (Store, Date) duple within the test set
* **Store** - a unique Id for each store
* **Sales** - the turnover for any given day (this is what you are predicting)
* **Customers** - the number of customers on a given day
* **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
* **StateHoliday** - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays.
* **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
* **StoreType** - differentiates between 4 different store models: a, b, c, d
* **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
* **CompetitionDistance** - distance in meters to the nearest competitor store



* **CompetitionOpenSince[Month/Year]** - gives the approximate year and month of the time the nearest competitor was opened
* **Promo** - indicates whether a store is running a promo on that day
* **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
* **Promo2Since[Year/Week]** - describes the year and calendar week when the store started participating in Promo2
* **PromoInterval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew.



# Feature Selection

* **Store**
* **School Holiday**

****

* **Promo**

****

* **Store Type**

****

* **Assortment**

****

* **Comp Open since month**

****

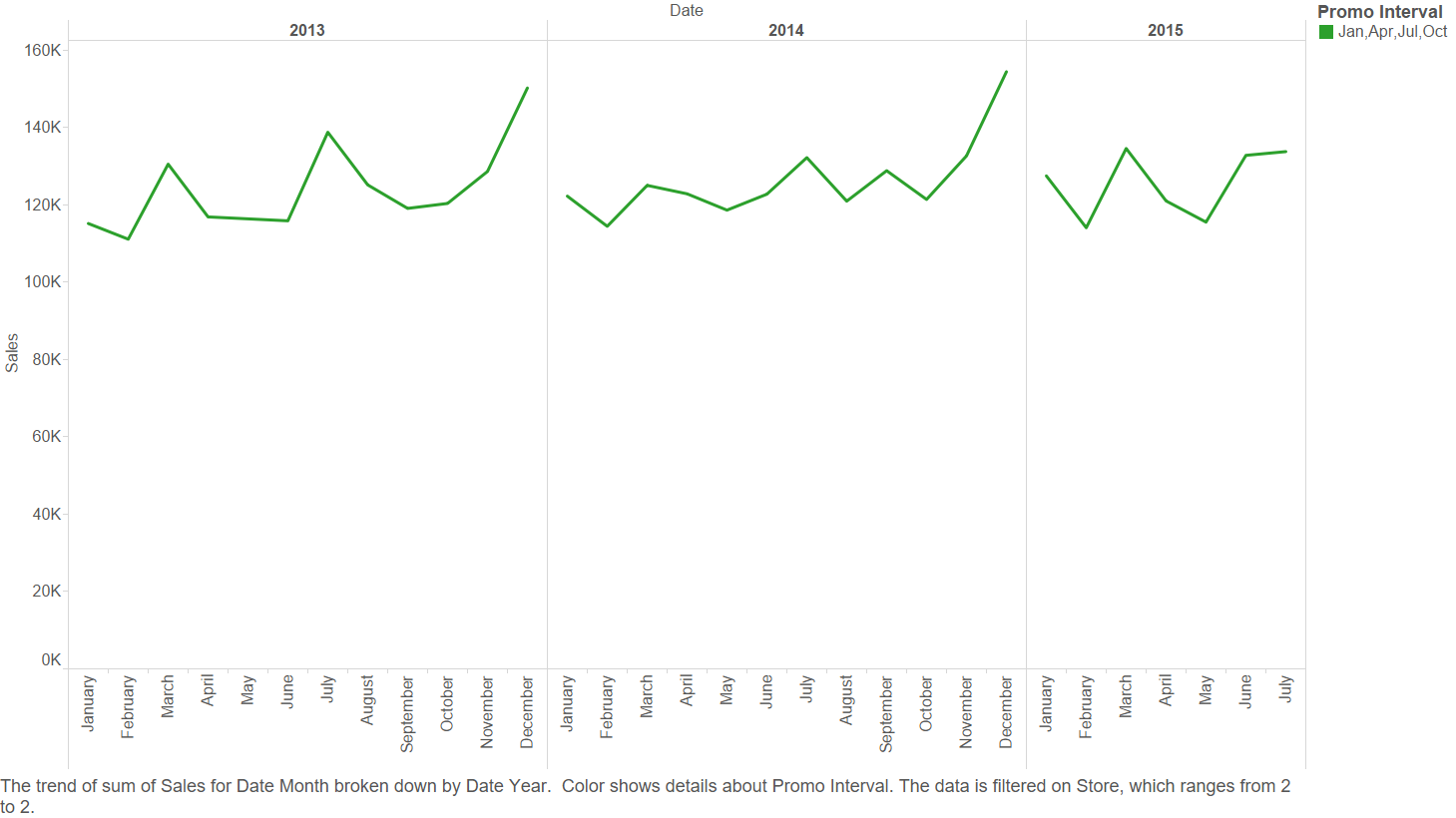
* **Comp distance**



# Feature Engineering

Linear Regression is a modeling technique that was used to model the relationship between a scalar dependent variable and an explanatory variable. In linear regression, unknown model parameters are being estimated from the available data using a linear predictor function.

* PromoMonth : We flag a store for all the months on which it had been giving out promotions



* CmpMsr: We made 2 assumptions for using this metric, they are
  1. An increase in competitor distance would mean increase in Sales of Rossmann stores
  2. An increase in the time a competitor has been present around a Rossmann store would mean a decrease in Sales of Rossmann Stores

The way the feature was engineered was :

1. Normalization of competitor distance
2. Competitor open since time calculated in terms of years
3. The two fields are multiplied
4. Result is normalized

* Splitting of Date: The sales data that is awailable is split as per the date



* Splitting of Month, Year: The sales are separated as per the month and year



* Splitting of Day of the Week: The sales data are split as per the day of the week



* Splitting of Week of the Year: The sales data are split week wise



# Main techniques

## Linerar Regression

Linear Regression is a modeling technique that was used to model the relationship between a scalar dependent variable and an explanatory variable. In linear regression, unknown model parameters are being estimated from the available data using a linear predictor function.

We used linear regression in our problem since it is the most basic and commonly used method for predictive analysis. Implementing the method was very easy and fast. We took into consideration the following factors to predict the sales of the stores for the next 6 weeks:

* Store
* School Holiday
* Competitor distance
* Year
* Month
* Promotion period
* cmpmsr

Different combinations were tried and tested to minimize the squared error. The available datasets were divided into three sets, each time two sets were used to predict the data for the third set and the predictions were checked against the actual dataset. Cross validation was used to minimize the square error.

## Random Forest

Random forest is a general technique of decision forest that has a learning process for regression, classification and other methods that require the construction of decision tree while training and then using the class that is the mode or the mean of individual trees.

A random forest is an ensemble of decision trees, it selects the rows and columns to form the trees. Random forest use averaging to find a natural balance. The reason we used the random forest method is because it captures the non-linear models well. We took into consideration the following factors to predict the sales of the stores for the next 6 weeks:

* Store
* School Holiday
* Competitor distance
* Competitor open since month
* Competitor open since year
* Year
* Month
* Special promotions
* Promotion period
* cmpmsr (engineered feature)

## Gradient Boosting

Gradient boosting is a powerful machine learning technique that is being widely used today. Ensemble methods are proving to be more efficient than other methods.

After working on a strong ensemble method like random forest, the next step was move away from averaging model and use some constructive ensemble strategy.

In gradient boost method, the learning method fits models consecutively and hence provides a very accurate estimation. Gradient boosting methods are known for their high customization and flexibility with error functions.

Visualization of the ensemble of trees becomes almost impossible due to the complexity involved in interpreting hundreds if not thousands of trees. The shrinkage parameter plays an important role in generalization of the model. A value lower than 0.1 is shown to produce a generalized model if there are sufficient number of rounds of boosting.

This effects the value of the leaves of the tree and not how the tree is built. The other parameters such as number of rounds of boosting, maximum depth of tree have to be chosen to avoid over-fitting. For implementing this, we use XGBoost which implements gradient boosted regression trees and linear model.

# problems Faced during implementation

## Linerar Regression

* Problem is not Independent and identically distributed random variables.
* It is time series
* Splitting data into training and cross validation has to be time based

## Random Forest

* Parameter Tuning
  + Increase number of trees without overfitting.
  + Decrease depth by increasing min\_samples\_leaf, min\_samples\_split
* Slow execution

## Gradient Boost

* Choosing number of boosting rounds, max depth of tree
* To avoid over fitting

# future trends

Sales prediction is a common and widely studied problem. Yet there are no specific models that are used universally. Regression is one of the commonly used methods though in this problem, gradient boosting proved to be most efficient.

The challenging goal would be to create a new model that takes into account the time series nature of data without explicit processing. This being a long term goal, short term goals on the project would include getting more external data to improve on efficiency.

As mentioned in the forums and the policy of kaggle which permits usage of external data, a better efficiency could be reached by having details such as weather. Another direct application could be to have a general prototype of all features that are useful in sales prediction.

##### References

[1] [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/](http://www.google.com/url?q=http%3A%2F%2Fwww.ncbi.nlm.nih.gov%2Fpmc%2Farticles%2FPMC3885826%2F&sa=D&sntz=1&usg=AFQjCNGBH2YERmJP8cGN2OS5sjypcW_r-A)

[2] [https://en.wikipedia.org/wiki/Gradient\_boosting](https://www.google.com/url?q=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FGradient_boosting&sa=D&sntz=1&usg=AFQjCNEZvbCDhqSQArPVt24WFkOfjOLuAg)

[3] [http://startup.ml/blog/xgboost](http://www.google.com/url?q=http%3A%2F%2Fstartup.ml%2Fblog%2Fxgboost&sa=D&sntz=1&usg=AFQjCNFzMALpijSPsCiQUhLhzFjtqrNNew)

[4] [https://www.kaggle.com/c/rossmann-store-sales](https://www.google.com/url?q=https%3A%2F%2Fwww.kaggle.com%2Fc%2Frossmann-store-sales&sa=D&sntz=1&usg=AFQjCNHxI9w0L32MsTs48Z_peEAvlvwdHQ)

[5] [https://www.kaggle.com/c/rossmann-store-sales/forums](https://www.google.com/url?q=https%3A%2F%2Fwww.kaggle.com%2Fc%2Frossmann-store-sales%2Fforums&sa=D&sntz=1&usg=AFQjCNFuYucxT8glRiXCP9byuoWvVyzlgA)

[6] [https://en.wikipedia.org/wiki/Linear\_regression](https://www.google.com/url?q=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FLinear_regression&sa=D&sntz=1&usg=AFQjCNF2nGgxMDKtImN3P8H_yFIubo178Q)

[7] [https://www.dataquest.io/course/kaggle-competitions](https://www.google.com/url?q=https%3A%2F%2Fwww.dataquest.io%2Fcourse%2Fkaggle-competitions&sa=D&sntz=1&usg=AFQjCNE2y0SBYNv1blrzVJ3Yc-GgfKI5tQ)