

Group 3 -Phoenix

Jatin Kakani

Mohamed Salihdeen

Supriya Salunkhe

Vigneshwaran Kannan

29th April, 2016

Rossmann Sales Forecasting

Table of Contents

**Rossmann Sales Forecast1**

**Objective2**

**Data2**

**Data Fields**………………………………………………………………………………………………………………………………….**3**

**Data Views**………………………………………………………………………………………………………………………………… **4**

**Hypothesis3**

Hypothesis 1**5**

Hypothesis 2………………………………………………………………………………………………..........................**6**

Hypothesis 3………………………………………………………………………………………………………………………..**7**

**Predictive Theorem**…………………………………………………………………………………………………………………… **8**

**Conclusion**………………………………………………………………………………………………………………………………**..11**

# Rossmann Sales Forecast

Rossmann operates over 3,000 drug stores in 7 European countries. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and sales itself. As of now, sales prediction at Rossmann’s stores are done by the store managers themselves without any assistance from statistical methods like the one that we’ve used in this project.

# Objective

The main objective of the project is to analyze the data and predict the sales of the stores in the near future thus recommending the board of directors of the organization whether or not to open a store in the future.

# Data

The data files which we are using for our project are listed below:

|  |  |
| --- | --- |
| **Data Files** | **Description** |
| train.csv | Historical data including sales |
| test.csv | Historical data excluding sales |
| store.csv | Supplement Information about the sales |

This dataset was posted in [www.kaggle.com](http://www.kaggle.com) as a part of competition. We made use of this dataset to predict the sales of the stores. Some of the stores in the dataset were temporarily closed for refurbishment. We have various measures in our dataset like Sales, Customers, and Competition Distance. Dataset did not have any missing values as well as outliers. As a part of data preprocessing, we selected 50 stores out of available 1115 stores and did a prediction on the sales.

# Data Fields

Most of the fields are self-explanatory. The following are descriptions for those that aren't.

|  |  |
| --- | --- |
| **Data Field Name** | **Description** |
| 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. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None |
| 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. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store |

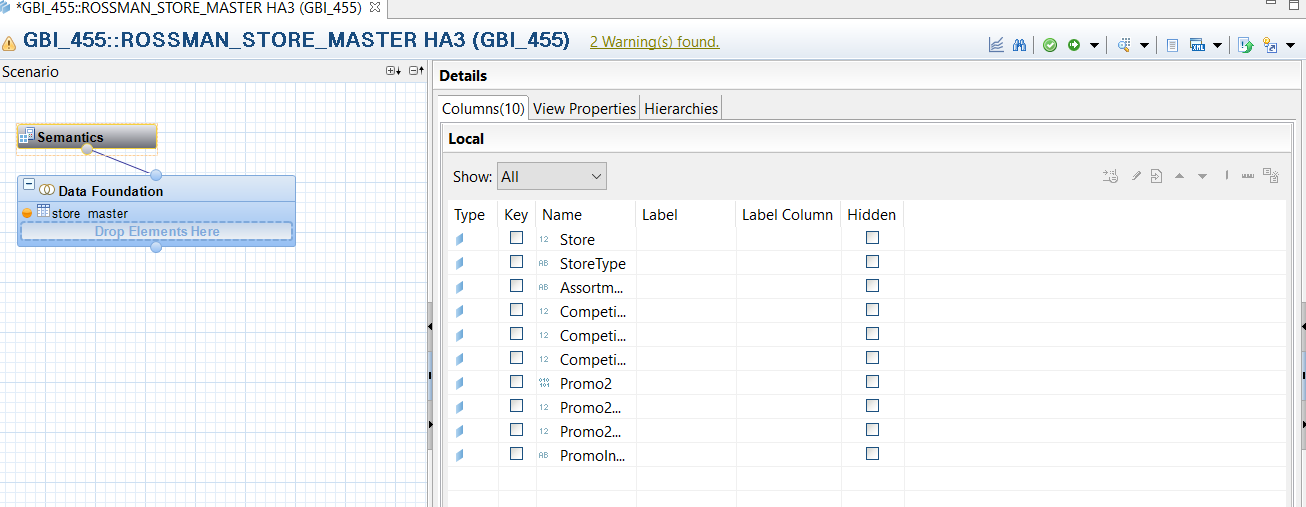
We have not considered the following data fields Promo2, Promo2Since[Year/Week] and PromoInterval

# Data Views

Two views attribute view and analytic view has been created using SAP HANA Modeler.

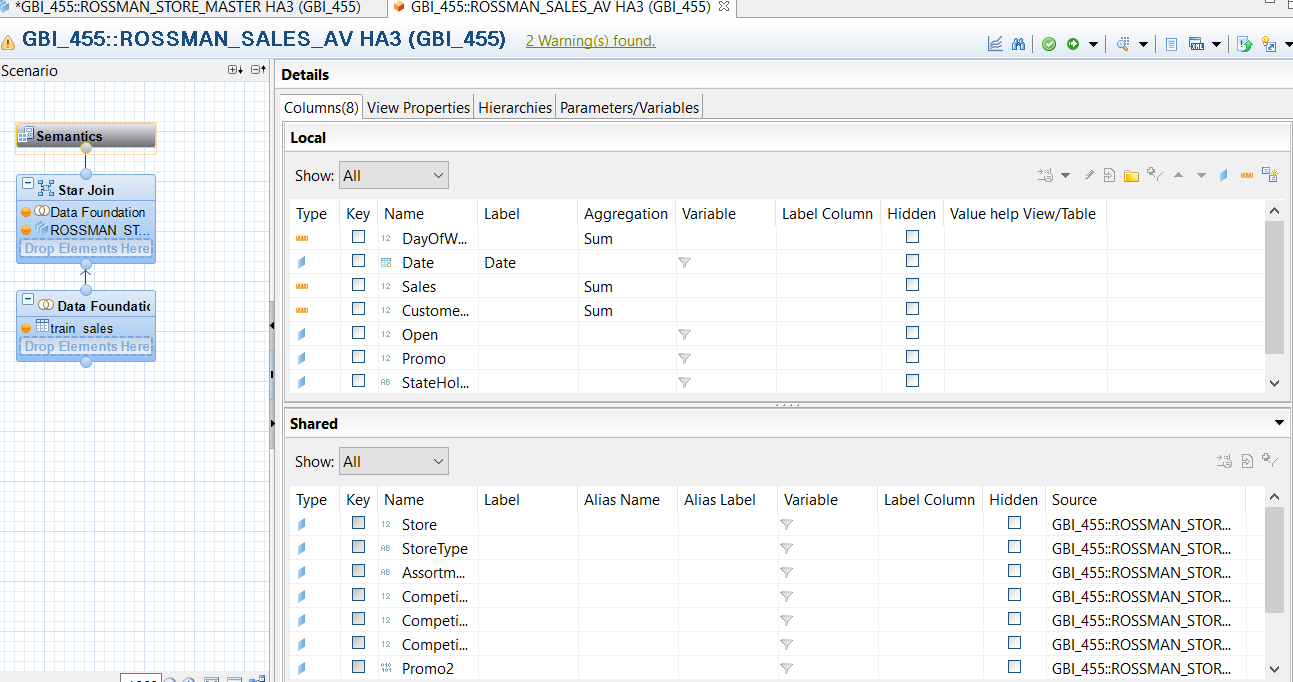
*Attribute view:*

Attribute view gives brief description about the master data used in the project. There are 10 data fields. Below is the screenshot.



*Analytic view:*

Analytic View is basically used to represents star models in HANA where the fact table is surrounded by different dimensions (master data).



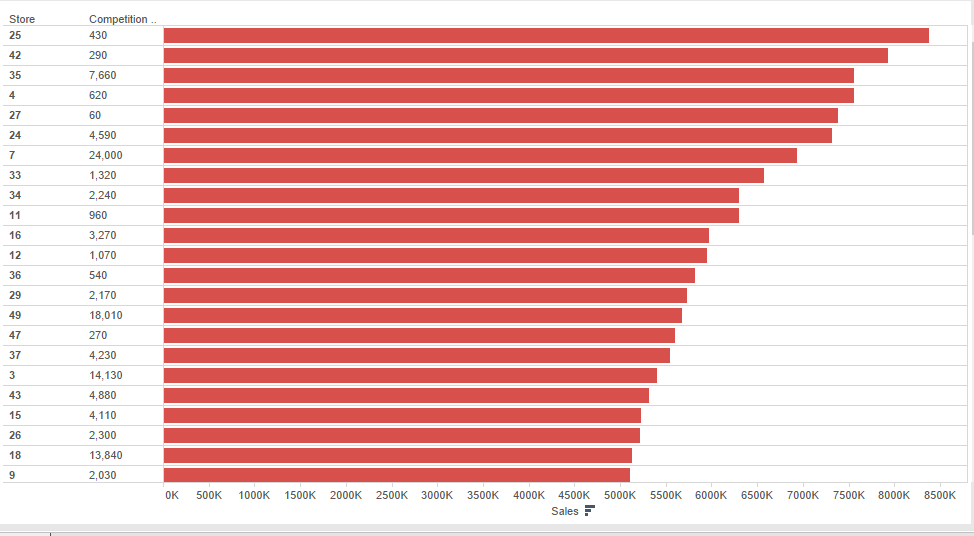
# Hypothesis

As a part of our detailed analysis, we came with four different hypothesis using different measures from the dataset. Few of the hypothesis were correct and few were proven wrong.

*Hypothesis 1:*

* If there is an increase in the competition distance, then there is an increase in the sales.

The graph below plots competition distance and sales for different stores.



*Conclusion:*

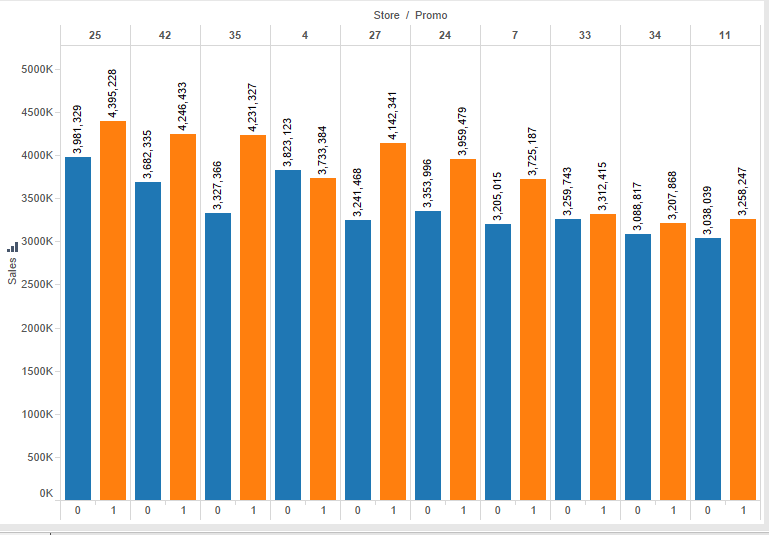
From the above graph we can depict that competition distance increases, it necessarily does not increase sales.

The graph is in decreasing order of sales and we can conclude that competition distance do not correlate with sales. Hence this hypothesis fails.

*Hypothesis 2:*

* When there is a promotional offer, the daily sales increases.

The graph below plots sales and promo for different stores.



*Conclusion:*

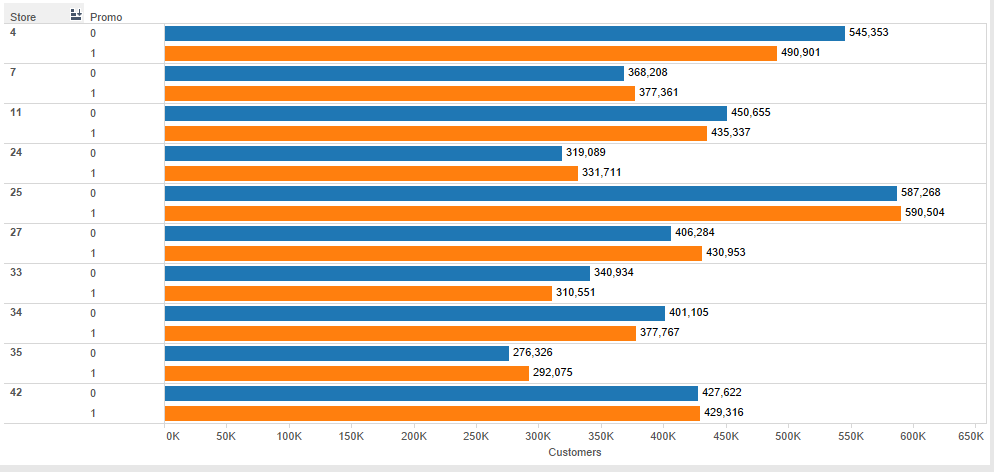
From the above graph we can depict that there is an increase in the sales when there is a promotions going on in a particular store.

We conclude that the sales increases when there is a promotional offer in a particular store. Hence this theorem succeeds.

*Hypothesis 3:*

* When there is a promotional offer, the number of customers that visit the store increases.

The graph below plots promotional offer and customers for different stores.



*Conclusion:*

From the above graph we can depict that there is no relationship between promotional offers and the number of customers visiting the store.

Hence this theorem fails.

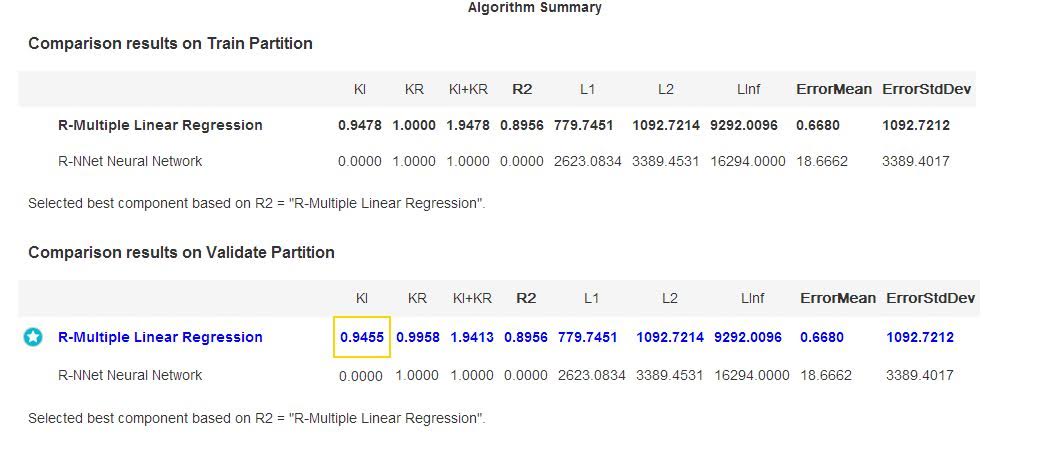
# PREDICTIVE THEOREM

The theorem aims at forecasting the sales for a set of Rossmann stores.

*Process:*

The below points highlight the process we carried out with the data during the prediction process:

* The dataset was partitioned into training, testing and validation datasets as per the following partitioning criterion:
  + Training: 50%
  + Validation: 30%
  + Testing: 20%
* We considered using Neural Networks model but after considering the correlation coefficient we decided to go with Multiple Linear Regression



*Regression Equation:*

**Y = -37.45 x1 + 6.968 x2 + 1077 x3 + 1146 x4 + 0.032 x5 + 37.10**

Y – Forecasted Sales

X1 – Day of week

X2 – Number of daily customers

X3 – Open flag

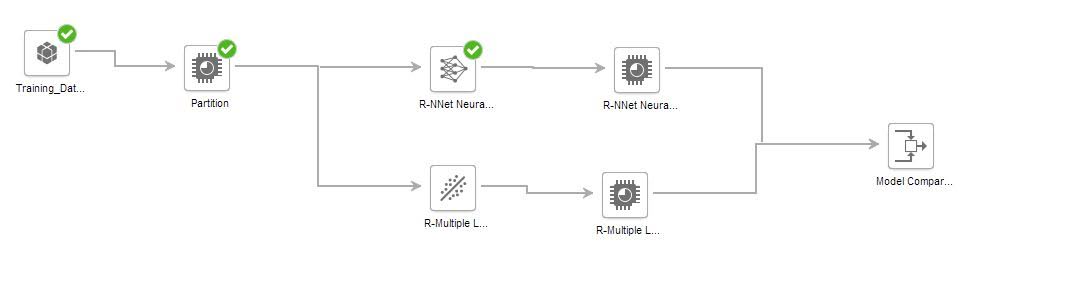
X4 – Promotion flag

X5 – Competition distance

Correlation Coefficient / R-square value: 0.8986

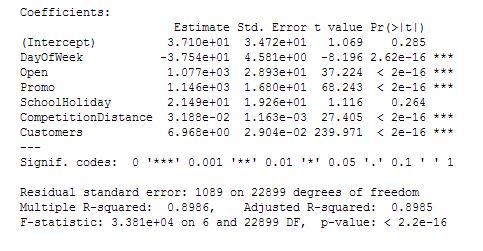
*Screenshots:*

*Predictive Model:*



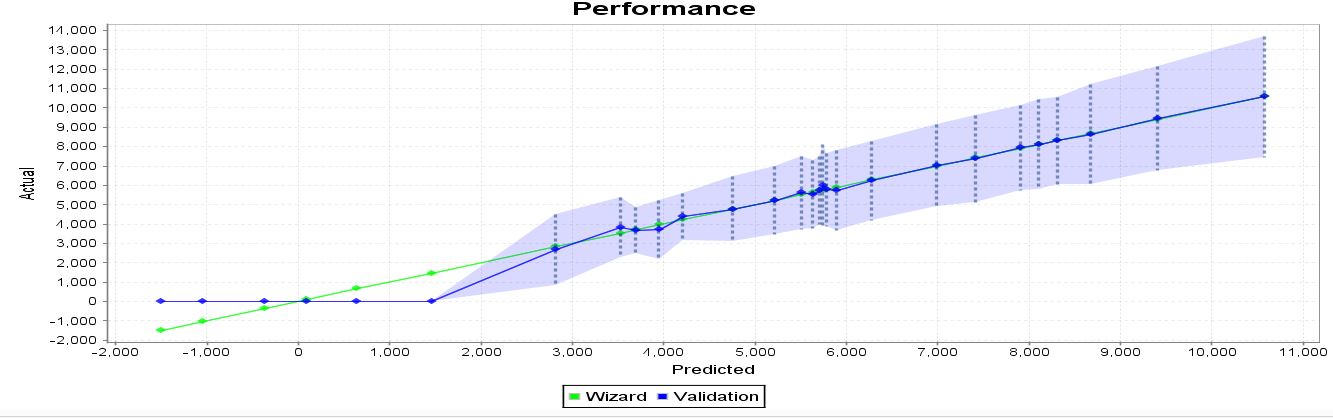
The above diagram shows the whole model that was used to forecast sales.

*Regression Table:*



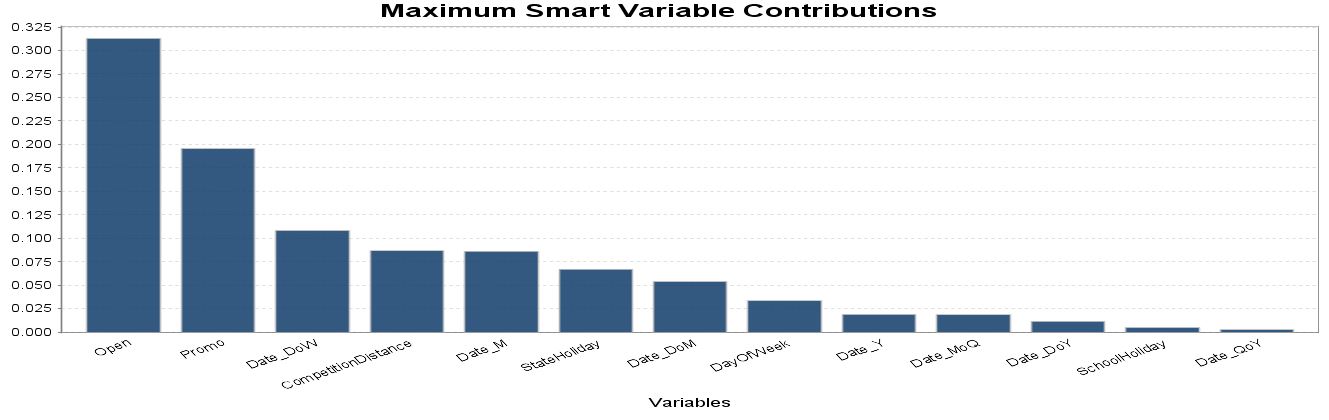
The regression table gives information about the variables that affect the daily sales of a store. The coefficients indicate the effect – positive or negative, on the sales.

*Model Performance (Actual v Predicted values):*



The above plot shows actual sales values and the predicted sales values. Both the lines align a lot similarily indicating that the model is valid.

*Variable Contributions:*



The above graph shows which variables contribute the most to the sales.

*Conclusion:*

The multiple regression model helps in forecasting the sales for Rossmann stores.  
It also explains why two of the hypotheses stated in this report failed and the impact of different attributes on the sales.