Trend Projection using Predictive Analytics

Seema L. Vandure
KLS Gogte Institute of Technology,
Udyambag, Belgaum
Karnataka, India

Manjula Ramannavar
KLS Gogte Institute of Technology,
Udyambag, Belgaum
Karnataka, India

Nandini S.Sidnal, Ph.D
KLE's College of Engg. & Technology,
Udyambag, Belgaum
Karnataka, India

ABSTRACT

With the growing use of social media networks, trends are being discussed and talked about everywhere. Trend Analysis is a skeletal mapping of expected changes or activities occurring in the societies, markets, organizations and the consumers who drive them. Past trends and patterns in the data can be studied and used, to make predictions for future. Regression is the commonly known technique to perform predictive analytics. In this system Linear Regression and SVM is analyzed for efficiency. Future sales trends are predicted using both the model and they are compared. Even impact of Google trends data on market sales is analyzed. Finally we conclude that search trends are useful in prediction of market sales where correlation is high and we also indicate that SVM is better to perform predictions.

General Terms

Data Mining, Predictive Analysis

Keywords

Predictive Analysis, Trend Projection, Linear Regression, Support Vector machines

1. INTRODUCTION

Day by day people are getting addicted to virtual world’s WEB life, i.e., social networking sites like Facebook, twitter or blogs etc. People are very eager to upload their life events – through pictures or through comments. This new lifestyle has become a common trend among people of all age groups. But in the midst of this venture, a new scope for intelligent analysis that is growing day by day is nothing else but our “DATA”. Data Analytics, a recent research trend deals with the above issue and captures meaningful insights from this data which can create value.

In the recent era, consumers have been remarkably quick to adopt trends and developments. Due to the changes occurring in social, political, and technological environments, opinion of public have been changing rapidly. This is why it is now crucial to identify and follow the early waves in the consumer ocean. Trend analysis is a structural mapping of expected changes in the behavior of societies, markets, and the consumers who drive them. Trends tend to develop within different time frames and on different levels. They can be short term, medium term or long term. Trend analysis gives companies the opportunity to innovate with less stress [1], [2]. Finding significant trends from large data sets has variety of applications.

Projecting trends help the businessman and economists to keep track of the latest happenings and also predict the trends for future which can help them to increase profit percentage. In this system, monthly retail sales data of US Census Bureau[1] is analyzed for 13 different NAICS retail trade categories. This data is analyzed for linear and SVM model and their performances are compared. Also the search trend data available on Google trends[2] is collected and mapped to these categories. Effect of Google search on market is analyzed and compared using linear and SVM model.

2. PROBLEM DEFINITION

Trends have become an important part in everybody life, and with increase in social networking, this feature has taken new turn into analysis phase. The objective of trend analysis is to determine either increasing or decreasing pattern in data which means respectively increasing or decreasing trend. Companies can analyze the past historical data to predict the trends and react accordingly so as to improve their sales figures and gain profit.

Predictive analytics helps us to predict or forecast the future events, strategies or policies based on past performance or experience. In this system, predictive analytics is used to predict the future trend using batch data. Google’s economist Hal Varian and Hyunyoung Choi have hypothesized that the search carried on Google in the form of query may have some correlation with the recent economic activity or market sales or trade records and thus it can be useful in predicting the subsequent data releases in future. Actual sales data and the query percent on Google are analyzed. Linear and SVM predictive model with and without the Google trends is build and effect of involving Google trend and accuracy of model is studied.

3. LITERATURE SURVEY

Companies store large amount of data regarding their customers, personal and business details, but the traditional data bases are not sufficient enough. Data mining techniques were used to gather past statistics and patterns [3].Data mining methods include techniques which evolve from artificial intelligence, statistics, machine learning, OLAP and so on. Classification, association, prediction, clustering are usually the common methods. The choice of what data mining techniques to apply at a given point in the knowledge discovery processes depends on the particular data mining task to be accomplished and on the data available for analysis. The requirements of tasks dedicate to the functions of mining and the detail characteristics of tasks influence the feasibility between mining methods and business problems [4].

Several software packages such as SAS and SPSS have existed to solve regression problems because statistic

[1] http://www.census.gov/retail/

[2] http://www.google.co.in/trends/explore#cmpt=q

techniques matured quite early in business area. Traditional prediction methods come from statistic area, for instance, linear regression and non-linear regression [4].

A number of studies have been conducted on different forms of social networks like Del.icio.us, Facebook, Flickr, Linkedln, Google, Wikipedia and Youtube etc. Sitaram et al. demonstrated how social media content like chatter from Twitter can be used to predict real-world outcomes of forecasting box-office revenues for movies [5][6].

Most researchers apply evolutionary computation (EC), Genetic programming(GP) to the analysis of equity trends and believe that there exist opportunities to identify, and take advantage of, patterns that indicate that the price of an equity or other financial instrument will rise or fall in the near future [7].The short-term forecasting models are mainly based on the parametric regression methods at present, which are included the early historical average models, time series models, and Neural Network models [8].

4. TECHNIQUES USED

4.1 Regression

Regression technique is supervised learning model and is often used to predict continuous variables which are usually numbers. In regression, we can find proper dependencies between variables. It is used to find out how best the variables are related. Regression helps in estimating relationship between two or more variables. Mainly regression analysis is used to understand relationship between dependent and independent variable, that is, what will be the effect on dependent variable if we change the independent variable. Regression analysis provides variety of models to perform analysis, for example linear regression, ordinary least squares, polynomial regression, generalized linear model etc. Mainly, regression can be Linear or Multivariate.

4.1.1 Linear Regression (LR)

Best line to fit two variables is searched so that using one variable we can predict the other. We can visualize it as functional dependencies between two variables, that is, how much the value of independent variable influences the value of dependent one.

4.1.2 Multivariate/Multiple Linear Regression

Here multiple variables and their correlations and dependencies are studied i.e. data is fit into multiple dimension surfaces to find out the interesting patterns using which we can predict for future.

4.2 Linear Regression

Linear regression is the method used for modeling the relationship between a scalar dependent variable labeled as Y and one or more explanatory (predictor) variables denoted as X. The case in which one explanatory or predictor variable is present called simple linear regression [9]. In linear regression, data is assumed to follow straight line relationship and unknown model parameters, that is, the response variable are estimated from the data using the predictor variables. Usually, linear regression is termed as a model in which the conditional mean or value of Y (called a response variable) for the known value of X called a predictor variable), is an affine function (a function having linear function along with constant variable and which contains a graph which has straight line) of X. i.e., for example, for modeling n data points it requires one independent or predictor variable: Xi, one dependent or response variable Yi and two parameters, α and β:

$$\mathbf{Y_i} = \alpha + \beta \mathbf{X_i} + \mathbf{e} \qquad \text{for } i = 1, 2 \ldots n \qquad (4.1)$$

where, the variance of Y is assumed to be constant, and α and β are regression coefficients specifying the Y intercept and slope of the line (Equation 4.1), respectively, e is an error term indicating that for most of the real world situations, the (X,Y) points are not organized exactly in a straight line. More precisely, e indicates the difference between actual value and the predicted value.

In general, model for multivariate linear regression can be given as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots.. + \beta_k x_k + \varepsilon \qquad (4.2)$$

Where,

β_0 is the intercept

β_1 is the parameter associated with x1 (slope parameter)

k represents the number of independent variables x_1-x_k independent variables

y dependent variable

ε error term

For a given data sample where Y={y_1,y_2,y_3,y_4,…y_n} and X={x_1,x_2,x_3,x_4,…x_n}, equation (4.2) can be modified as:

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots\ldots.. + \beta_k x_{nk} + \varepsilon_n \qquad (4.3)$$

Equation (4.3) can be written in the matrix form as,

$$Y = X\beta + \varepsilon \qquad (4.4)$$

where,

$\beta = [\beta_0\, \beta_1 \beta_2 \ldots\ldots\ldots\ldots\ldots\ldots \beta_k]$

$\varepsilon = [\varepsilon_1\, \varepsilon_2 \ldots\ldots\ldots\ldots\ldots \varepsilon_n]$

The parameter represented by β in (4.4) is calculated by using least square estimates so that sum of squared of error is minimized. To compute the coefficient estimates Quadratic Decomposition (QR) method is used. While computing least squares, we calculate the coefficients β_0, β_1, β_2, ...β_k so that the sum of squared errors i.e. SSE is minimum (4.5) [10].

$$SSE = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \qquad (4.5)$$

4.3 Support Vector Machine

Support Vector Machines (SVM) has its root in decision planes which are used to define decision boundaries. A decision plane separates a set of objects as per the different class memberships. Formally, it can be stated as, it builds a hyperplane. It can be a single hyperplane or set of hyperplanes in a high-dimensional space, which can be used for different tasks like classification or regression. Hyperplane is said to have achieved a good separation if the distance between the nearest training data point is large (so-called functional margin), which means larger the margin the lower is the error of the model.

SVM is mainly used for classification task but can also be used for regression analysis if the predictor variables are continuous values. Rough sketch of SVM can be given as follows [11]:

- Class separation: Hyperplane must separate two classes by increasing the distance between the nearest points of class the points which are lying on the boundaries are called support vectors, and the center of the margin is called the optimal separating hyperplane.
- Overlapping classes: Data points which are lying on the wrong side of the discriminant margin need to be weighted down so as to reduce their effect or influence (soft margin).
- Nonlinearity: In practical situations, data may follow nonlinearity in such cases it is difficult to find a linear separator. In such cases either data points are reduced or projected into an (usually) higher-dimensional space, which converts data points in such a way that now it becomes linearly separable This can be done using kernel trick.

In general, SVMs belongs to category of kernel methods. Kernel method used is an algorithm which is dependent on the data only through dot-products [12] [13]. In such cases, the dot product can be computed using kernel function (in high dimensional feature space). Kernel function used has two advantages:

- It can create non-linear decision boundaries.
- User can apply regressing model to data which do not have fixed dimensional vector space

4.4 Error Measures

In order to find which model is good and also to check if trends data is helpful in predictions, we will find the prediction error using various available error measurements methods, such as, MSE, RMSE, MAE and MAPE. The forecast error is mainly defined as difference between actual value and predicted value. If y_i is actual value and $\hat{y}_i$ is predicted vale then the residual or error term is given as:

$$e = y_i - \hat{y}_i$$

where, i indicates the number of data points. While forecasting, error measures play an important role, it helps to attune the forecasting models. Refining the forecasting model as per error measures helps in increasing the accuracy of predictions, which is the main goal of predictive analytics.

Definitions for different error measures are given below:

- *Mean Squared Error* (MSE): In this case prediction errors or residuals are squared and then there average is calculated this is known as mean squared error.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  where n is the total number of data points.
- *Root Mean Squared Error* (RMSE): Since the errors were squared to calculate MSE we will take root to get more clear understanding of error. So, the root of average of sum of squared error is known as root mean squared error.

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

- *Mean Absolute Error* (MAE): The average of the absolute values of the prediction errors or residuals is known as the mean absolute error.

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

- *Mean Absolute Percent Error* (MAPE): The average of percent change in the absolute values of the forecasting errors (where percent change means division of prediction error by actual data) is known as the mean absolute percent error.

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

Model having minimum prediction errors are good for prediction cause less the error more will be the accuracy. In the further sections results of experiments will be analyzed and discussed using RMSE and MAPE as error measure.

5. System Design

5.1 Work Process of the System

Proposed methodology for this system is composed of five steps and step 3 and step 4 are repeated till accurate model is obtained (Figure 1).

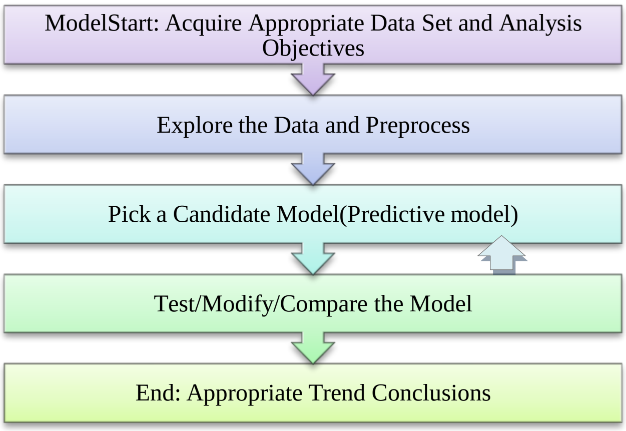
ModelStart: Acquire Appropriate Data Set and Analysis Objectives
Explore the Data and Preprocess
Pick a Candidate Model(Predictive model)
Test/Modify/Compare the Model
End: Appropriate Trend Conclusions

Fig 1: Proposed methodology and System design

5.1.1 Data Acquisition

Data is being collected from Google Trends in the form of comma separated value (.csv) file representing the search interest pattern of people. This data set contains search interest pattern recorded from 2004 till present. Google Trends provides an index of the volume of Google queries by geographic location and category. In this system we also consider US trends data for retail sector categories [14].

5.1.2 Explore data

While exploring data we can use data narrative, Scatter plots, Line and histogram plots Summary statistics Correlations among all variables Probability plots/data distribution to properly visualize and analyze data.

Table 1. Retail sales mapped with Google categories

		NAICS Sectors		Google Categories
S. No.	**ID**	**TITLE**	**ID**	**TITLE**
1	441	Motor vehicle and parts dealers	47	Automotive
2	442	Furniture and home furnishings stores	11	Home & Garden
3	443	Electronics and appliance stores	5	Computers & Electronics
4	444	Building mat., garden equip. & supplies dealers	12-48	Construction & Maintenance
5	445	Food and beverage stores	71	Food & Drink
6	446	Health and personal care stores	45	Health
7	447	Gasoline stations	12-233	Energy & Utilities
8	448	Clothing and clothing access. Stores	18-68	Apparel
9	451	Sporting goods, hobby, book, and music stores	20-263	Sporting Goods
10	452	General merchandise stores	18-73	Mass Merchants & Department Stores
11	453	Miscellaneous store retailers	18	Shopping
12	454	Nonstore retailers	18-531	Shopping Portals & Search Engines
13	722	Food services and drinking places	71	Food & Drink

5.1.3 Model

Supervised learning is the process of creating predictive models using a set of historical data that contains the results which we are trying to predict. In this system different models are analyzed for their accuracy and the model with minimal error rate is used for predicting the sales using trends data.

Working of proposed methodology is as follows:

5.1.3.1 Input data

It is the data collected from the Google trends and actual sales data of US retail sector. Attributes involved are Date specifying the duration over which search rates and sales were collected and volume indicating the value of search rates and sales amounts.

5.1.3.2 Training set and Test set:

Data is then divided into training and test set. The preferred ratio is 70:30, 70% of data is used for model training. Once the model is trained we use the remaining 30% of test data to improve accuracy.

5.1.3.3 Model Trainer:

Model is trained using training data. Linear model and SVM model both are trained for the training data. We even calculate the root mean squared error (RMSE) and mean absolute error (MAE) to find the how far our prediction does vary.

5.1.3.4 Check Accuracy:

The predictive accuracy of the model is estimated, the accuracy of a model on a given test set is the percentage of test set samples that are correctly predicted by the model

5.1.3.5 Predictions:

After the model has been processed, the results are stored as a set of statistics together with the linear regression formula, which you can use to compute future trends.

5.1.4 Test /Modify /Compare model

Model generated in earlier step is been tested for different data samples and modified as per their attributes and requirements. Even we can compare our predictions with other model to find the best fitting model.

5.1.5 Trend Conclusion

Based on the results of accurate fitting model the final conclusion is provided. This conclusion will provide us with knowledge that whether sales for particular sector shows a increasing or decreasing trend and what sequence will they follow in future..

6. Model Creation

Initially data is collected and integrated for different categories available in retail sales and Google trends.

Table 1 represents top level NAICS categories and their related subcategories in Google Trends [14]. Data is further divided into training data and test data. Models are initially trained using trained data and then tested for accuracy using test data. Simple model for US market sales will be written as follows:

$$\text{Model:} \quad Y_i \sim X_i + e \qquad (6.1)$$

Here Y_i indicates sales and X_i is the time values for i number of data points. Moving towards model building, the method used in R to perform linear modeling is *lm* and it uses least square method to estimate coefficients. Least square estimates can be computed using estimates Quadratic Decomposition (QR) method.

Working algorithm for linear model can be given as follows [15]:

Step 1: Fix the values α and β given in (4.1)

Step 2: For given predictor variable (i.e., X), make a guess for associated response variable Y.

Step 3: Error term is calculated by subtracting predicted weight from true weight.

Step 4: Error term calculated in step 3 is squared.

Step 5: Sum the squared error (SSE) calculated in step 4 for the given data points.

Further same data is analyzed with SVM model. As said earlier it generates optimal separating hyperplane which separates two classes. SVM models data to identify decision boundary using the kernel trick. For modeling data using SVM we follow following procedure.

- Initially build a model for SVM using linear kernel.
- Check if the models performance can be improved using non linear kernel.

Kernel level functions are used to transform data into high dimensional space, through which decision boundary can be easily described. There is significant effect of Kernel function on the decision boundary. If polynomial kernel or Gaussian kernel is selected then parameters available with these kernels like degree for polynomial and width for gaussian can affect the resulting model [12]. This kernel function can take four values linear, radial, polynomial, and sigmoid.

Now the correlation between actual and trend data is calculated. The value of correlation function can be 1, -1 or 0 [15]. Positive value (nearer to 1) indicates that there exist positive correlation between the market sales and the queries data. Negative value (nearer to -1) there is negative correlation between the two, and value nearer to 0 indicates that the two variables are not correlated. Threshold for correlation has been assigned value as; if it is above 0.3 then it is 'better' correlation, above 0.5 has 'good' and above 0.7 has strong correlation.

7. RESULT ANALYSIS

Table 2 displays the values obtained by different error measures, for different models on the testing data for category 8. First column lists what kind of model is used, rest all columns indicate MSE, RMSE, MAE, MAPE values.

Table 2. Prediction errors for test data

NAME	MSE	RMSE	MAE	MAPE
Linear without Trends	1287127	1134.516	928.0874	4.390212
SVM without Trends	750821.6	866.4996	691.1052	3.329109
Linear with Trends	516339.2	718.5675	586.6126	2.915033
SVM with Trends	450477.7	671.1764	539.8637	2.702466

In comparison of the linear with trends data and without trends data it is observed that the linear model with trends data performs better having RMSE value 36.7% less and MAPE value 33.6% less. Similarly when SVM model is compared using with trends data and without trends data, the RMSE and MAPE values for SVM model with trends data is 22.54% and 18.82% less respectively.

By the above experimentation it is observed that the models perform best when trends data is included. Further comparing, the linear and SVM model with trends, the RMSE value is 6.6% less and the MAPE value is 7.3% less for the SVM model which includes trends data as compared to linear model.

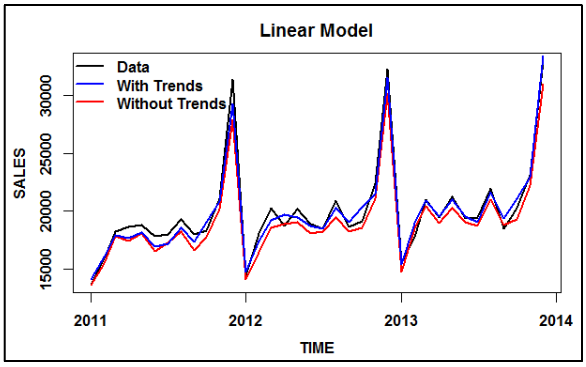

Fig 2: Out of sample predictions using linear with trends and linear without trends

Table 3 Error measures for linear model with and without trends

NAME	RMSE	MAPE
Linear without Trends	1134.516	4.390212
Linear with Trends	718.5675	2.915033

Figure 2 represents the values forecasted by linear model when trends data is included and excluded. The real value for the test data is represented in black color while the values predicted by linear model when trends data is included is shown in blue color and when the trends data is excluded, the predicted values are shown in red color. Table 3 displays the RMSE and MAPE values where it can be observed that when data with trends is used then model has smaller errors (RMSE - 36.7% and MAPE- 33.6% small).

Similar graph is obtained when SVM model with trends and without trends is used as shown in Figure 3. Actual data is displayed in black color; data with trends is in blue color, while data without trends is in red color. From Table 4 we can observe that error measure for data with trends is smaller as compared to without trends. For entire forecast RMSE is 22.54% small and MAPE is 18.82% small.

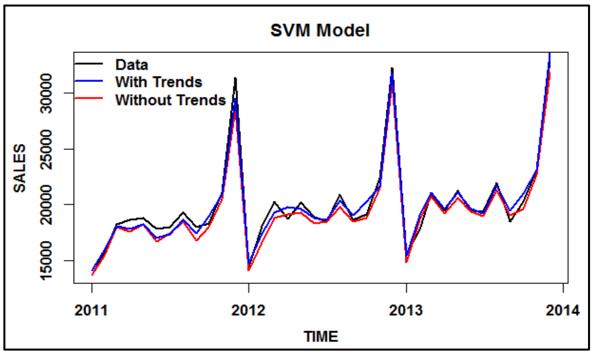

Fig 3: Out of sample predictions using SVM with trends and SVM without trends

Table 4 Error measures for SVM model with and without trends

NAME	RMSE	MAPE
SVM without Trends	866.4996	3.329109
SVM with Trends	671.1764	2.702466

The values predicted by both linear and SVM model when trends data is included is shown in Figure 4. It can be observed in this plot that both these models have approximately predicted similar values.

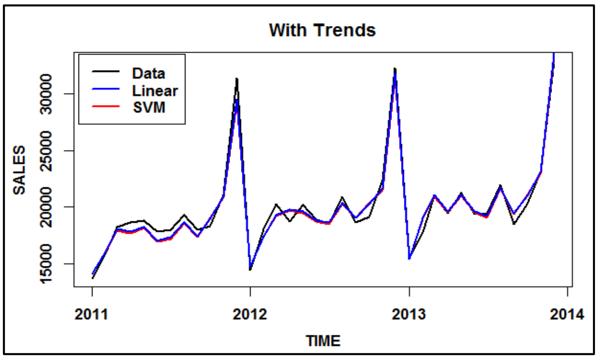

Fig 4: Out of sample predictions using linear with trends and SVM with trends

Table 5 Error measures for linear model and SVM model with trends

NAME	RMSE	MAPE
Linear with Trends	718.5675	2.915033
SVM with Trends	671.1764	2.702466

Table 5, displays the RMSE and MAPE measures for both linear and SVM model; when both of them uses trends data for prediction. In this case SVM has performed better with RMSE value 6.6% less and MAPE value 7.3% less as compared to linear.

Figure 5 shows a barplot of the prediction errors, for the four models. Yellow color indicates prediction error when trends were included in linear model and red color indicates errors when trends data was not considered by linear model. For SVM model, error are given by blue color when trends not considered and by green color when considered. In both linear and SVM models for most of the months when trends data is considered the prediction errors are low.

8. CONCLUSION AND FUTURE WORK

A track of early signals or trends helps organizations to be prepared for any events that may occur in future. Proper analysis will yield proper outcomes, and while performing analysis, web search data has always turned out to be useful. The idea of using Google’s search query data for predictive analytics turned out to be a successful indicator for accurate predictions. From the experimental results, we can infer that SVM model is better as it was observed that prediction errors are small in case of SVM compared to linear model. Also, while performing prediction, if Google trends data were added then prediction errors were lower for most of the months, as compared to predicting without Google Trends query index.

Based on our work as well as the current state-of-the-art, the horizon can be expanded. Different other sub categories can also be included in combination with the above mentioned categories to check their impact. Also, the model can be tried with different tuning parameters to obtain more accurate result.

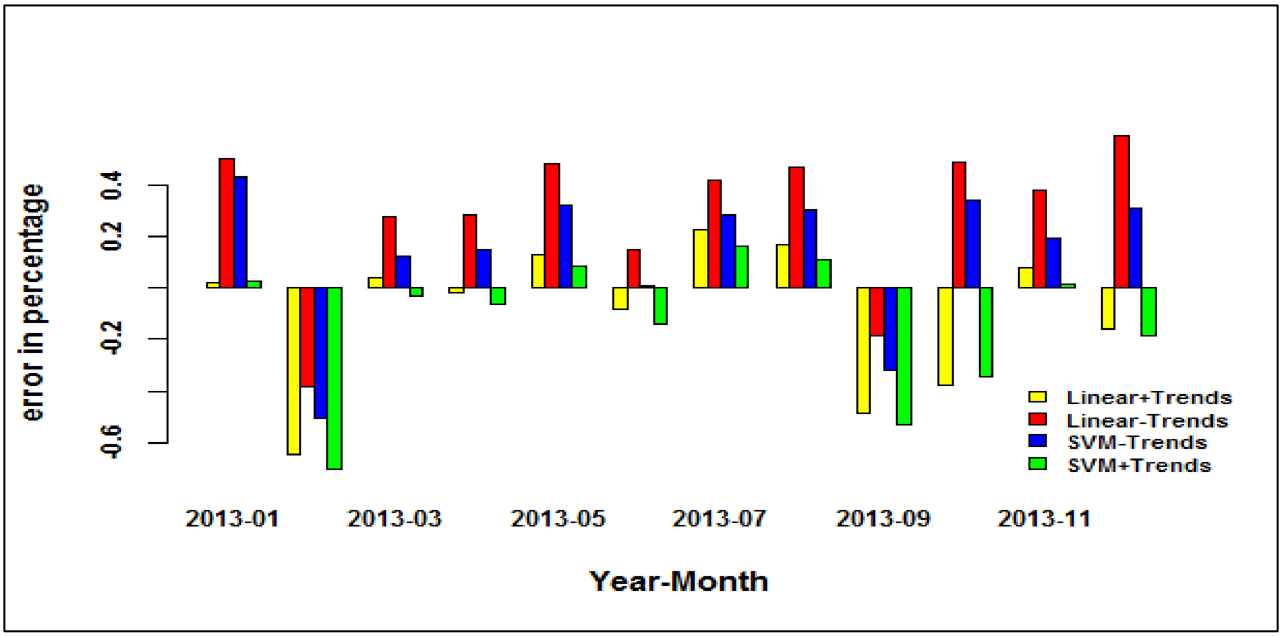

Fig 5: Prediction error barplot for Linear and SVM model

9. REFERENCES

[1] Justien Marseille and Ilan Roos, "Trend Analysis: An Approach for Companies that Listen," Design Management Review, pp. 68-72, 2005.

[2] Sreenivas Gollapudi and D. Sivakumar, "Framework and Algorithms for Trend Analysis in Massive Temporal Data Sets," ACM, 2004.

[3] Dirk Van den Poel, , Dirk Thorleuchterb Jeroen D'Haena, "Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique," Expert Systems with Applications 40, pp. 2007-2012, 2013.

[4] Jia-Lang Seng and T.C. Chenb, "An analytic approach to select data mining for business decision," Expert Systems with Applications 37, pp. 8042-8057, 2010.

[5] Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, Benyuan Liu Harshavardhan Achrekar, "Predicting Flu Trends using Twitter Data," The First International Workshop on Cyber-Physical Networking Systems,IEEE, pp. 702-707, 2011.

[6] Fabian Abel, Geert-Jan Houben, Ke Tao Qi Gao, "Interweaving Trend and User Modeling for Personalized News Recommendation," IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 100-103, 2011.

[7] Garnett Wilson, "Using Sector Information with Linear Genetic Programming for Intraday Equity Price Trend Analysis," World Congress on Computational Intelligence, IEEE, 2012.

[8] Yuan-Hui Wang Zheng-Wu Yuan, "Research on K Nearest Neighbor Non-parametric Regression Algorithm Based on KD-Tree and Clustering Analysis," in Fourth International Conference on Computational and Information Sciences,IEEE, 2012, pp. 298-301.

[9] http://en.wikipedia.org/wiki/Regression_analysis.

[10] Rajendra Banjade and Suraj Maharjan, "Product Recommendations using Linear Predictive Modeling," IEEE, 2011.

[11] David Meyer, "Support Vector Machines The Interface to libsvm in package e1071," Jan 2014.

[12] Asa Ben-Hur and Jason Weston, A User's Guide to Support Vector Machines.

[13] David Meyer and Kurt Hornik Alexandros Karatzoglou, "Support Vector Machines in R," Journal of Statistical Software, vol. 15, no. 9, April 2006.

[14] Hal Varian Hyunyoung Choi, "Predicting the Present with Google Trends," Google Inc., December 2011.

[15] Drew Conway and John Myles White, Machine Learning for Hackers, 1st ed., Julie Steele, Ed.: O'Rielly, 2012.