



SALES PREDICTION WITH MACHINE LEARNING

¹Dr. VVAS LASKHMI, ²PATIBANDLA VASANTHA LAKSHMI, ³KUMMETHA TRIVENI, ⁴KOTHAMASU MOUNIKA, ⁵KOTHA GOPICHAND

^{1,2,3,4,5}Assistant professors, Department of CSE in Narasaraopet Institute Of Technology

ABSTRACT

In this project, we study the usage of machine-learning models for the sales predictive analytics. The main goal of this project consider main approaches and cases studies of using machine learning for sales forecasting. The effect of Machine-learning generalization has been considered. This effect can be used to make sales predictions when there is small amount of historical data for specific sales time series in the case when a new product or store is launched. Stacking approach for building regression of single models has been studied. The results that using stacking techniques, we can improve performance of predictive models for sales time series forecasting. This is the age of the internet where the amount of data being generated is huge that man alone is not able to process through the data. So many machine learning techniques hence have been discovered for this purpose. In this project, we are trying to predict the sales of a retail to store using different machine learning techniques and trying to determine the best of algorithms suited to our particular problem statement. We have implemented the normal regression techniques and as well as boosting techniques in our approach and have found that the boosting algorithms have better results than the that of regular regression algorithm. Sales prediction is current numerotrend in which all the business companies thrive and also organization or concern in determining the future goals for it and its plan and procedure to achieve it.

1. INTRODUCTION :

Sales forecasting can be defined as the prediction of upcoming sales based on the past sales occurred. Sales forecasting is of paramount importance for companies which are entering new markets or are adding new services, products or which are experiencing high growth. The main reason a company does a forecast is to balance marketing resources and sales against supply capacity planning. Forecasting can help in answering some expository queries like “Do we have the right mix of price, promotion, and marketing in place to drive demand?”, “Do we have enough salespeople to get the volume of orders we have budgeted?”, “Do we have the essential demand-side resources in place?” and for these reasons, many of the companies allocate significant financial and human resources to perform this task genuinely, which requires large investment. Manufactures organizations and business houses require an accurate and reliable forecast of sales data so that they don't suffer from losses due to wrong or inaccurate

prediction by the model. Companies mainly use sales forecasting to determine two things. First, to determine the current demand level of the service or product in the market. Second, to determine the future demand for a company's services or products. Forecasting can be used to predict sales revenue at product level, or at an individual business level, or at company level. In this project we have concentrated on product level sales forecasting. Future sales plan aids in optimal utilization of facility, scheduling, conveyance and effective control of inventory. These, in turn, result in enhancement of clients' satisfaction and also decrease in production cost. In the recent past, many investigations addressing the problem of sales forecasting have been reported. Sales forecasts affect a company's marketing plan directly. The marketing department is responsible for how clients and their customers interprets its services and products and compare it against its competitors and use the sales forecast to assess how marketing spending can increase sales and channel demand. It is important to



develop effective sales forecasting models in order to generate accurate and robust forecasting results. In the business and economic environment, it is very important to accurately predict various kinds of economic variables such as Past Economic Performance, Current Global Conditions, Current Industry Conditions, Rate of Inflation, Internal Organizational Changes, Marketing Efforts, Seasonal Demands, etc. to develop proper strategies. On the contrary, inaccurate forecasts may lead to inventory shortage, unsatisfied customer demands, and product backlogs. Due to these reasons, utmost importance is given to develop productive models in order to generate robust and accurate results. In this paper we will be considering a variety of forecasting methods such as Multiple Regression, Polynomial Regression, Ridge Regression, Lasso Regression etc. along with various boosting algorithm like AdaBoost, Gradient Tree Boosting so as to get the maximum accuracy. Multiple Regression is a statistical tool used to predict the output which is dependent on several other independent predictor or variable. It combines multiple factors to access how and to what extent they affect a certain outcome. Polynomial Regression is an extension of simple linear regression. Where the model finds a non linear relationship between the independent variable x and the dependent variable y . Here the model usually fits the variable y as the n th degree of variable x . Ridge Regression is a way to create a model when the predictor variable has multi-collinearity. Lasso is a regression method that performs both regularization and variable selection in order to enhance the prediction accuracy. The AdaBoost is a boosting algorithm which is mainly used for improving the performance of the models. It utilizes the output of the other weak learners and combines the outputs of those algorithms into a weighted sum and finally arriving to the output. AdaBoost can significantly improve the learning accuracy no matter whether applied to manual data or real data. The elastic net method overcomes the limitations of the Lasso.

Gradient Tree Boosting is also a boosting algorithm which creates a model in the form of a group of weak predictors. In this paper a wide range of forecasting methods are discussed because the combination of multiple forecasts can be used to increase the forecast accuracy.

2. LITERATURE SURVEY

Fantazzini, Z. Toktamysova, Forecasting German car sales using Google data and multivariate models, Int. J. Production Economics 170 (2015) 97-135

Long-term forecasts are of key importance for the car industry due to the lengthy period of time required for the development and production processes. With this in mind, this paper proposes new multivariate models to forecast monthly car sales data using economic variables and Google online search data. An out-of-sample forecasting comparison with forecast horizons up to 2 years ahead was implemented using the monthly sales of ten car brands in Germany for the period from 2001M1 to 2014M6. Models including Google search data statistically outperformed the competing models for most of the car brands and forecast horizons. These results also hold after several robustness checks which consider nonlinear models, different out-of-sample forecasts, directional accuracy, the variability of Google data and additional car brands

Zone-Ching Lin, Wen-Jang Wu, "Multiple LinearRegression Analysis of the Overlay Accuracy Model Zone", IEEE Trans. on Semiconductor Manufacturing, vol. 12, no. 2, pp. 229 – 237, May 1999.

The capacity to foresee information precisely is critical and significant in spaces, for example, stocks, deals, climate or it can likewise be utilized in Marketing Sectors. We have Presented the examination and executed a few characterization calculations utilized on deals information, comprising of week by week retail deals numbers from various divisions in Walmart retail outlets. The models executed for forecasts are KNN (k-closest neighbor), Random Forest, Extra Trees Regressor and

SVM (Support Vector Machine). A near investigation of these calculations is performed to show the best calculation and utilizing this Algorithms we can get the best Results.

O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, "Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis", Int. Journal on Mathematical Theory and Modeling, vol. 2, no. 2, pp. 14 – 23, 2012.

Regression analysis is used across business fields for tasks as diverse as systematic risk estimation, production and operations management, and statistical inference. This paper presents the cubic polynomial least square regression as a robust alternative method of making cost prediction in business rather than the usual linear regression. The study reveals that polynomial regression is a better alternative with a very high coefficient of determination.

3. SYSTEM ANALYSIS

3.1: Existing System

Sales forecasting is a sophisticated problem and is influenced by external as well as internal factors and there are two major drawbacks to the statistical approach as told by A. S. Weigend A hybrid seasonal quantile regression approach and (ARIMA) Auto-Regressive Integrated Moving Average approach for daily food sales forecasting were proposed by N. S. Arunraj and also found that the performance of the individual model was comparatively lower than the hybrid model. D. Fantazzini proposed new multivariate models which is used in forecasting the car sales in Germany using Google data. In his paper, he stated that long-term forecast is of at most value for the car industry due to the lengthy period of time required for production and development process. E. Hadavandi used an integration of Genetic Fuzzy Systems (GFS) and data clustering for the sales forecasting of the printed circuit board. In their paper by using K-means clustering created K clusters of all the

records of the data. Then, all the clusters were fed into independent Genetic Fuzzy Systems (GFS) with the ability of database tuning and rule-based extraction. Recognized work in the field of sale forecasting was carried out by P.A. Castillo, They performed sales forecasting on new published books in an editorial business management environment by applying computational methods. (ANN) Artificial Neural Networks are also used in the sales forecasting field. Fuzzy Neural Networks have been introduced with an objective to improve the prediction performance and also Radial Basis Function Neural Network (RBFN) is assumed to have a great potential for the prediction of sales The literature in this field shows that not much work has been done in swarm intelligence technique in effectively training the prediction models. The Genetic Algorithm (GA) is a potential candidate for training the ANN models. There are a lot many works in this field which has helped the organization to predict the future profit what they can make by investing at the proper place at right time. This paper is another contribution to this field.

3.1.1 Disadvantages

- No accurate results
- Loss the data base.

3.2: Proposed System

Having the sales data of the retail store, the proposed work suggests the following various steps for predicting the sales of different categories available. The architectural diagram for the proposed algorithm . The various steps involved are explained hereunder.

A. Hypothesis Generation This step is of primal importance in the process of data analysis. In this step various hypotheses are generated by analyzing the problem statement

B. Data Exploration When we consider a business problem, we try to achieve more accuracy by changing and implementing different models. The first step in Data exploration is to look into the dataset and to discover the information regarding the

available and the hypothesized data. The dataset under consideration has the following features as the variables.

C. Data Pre-processing This step typically imputes the missing values and handles the outliers present in the dataset. Missing values are found in Item_Weight and Outlet_Size columns of the dataset. Item_Weight is the numerical variable and Outlet_Size is the categorical variable.

D. Feature Engineering During the data exploration step, we identified a few of the nuances in the data. This step deals with those nuances and also is used for creating new variables out of existing variables so that our data will be ready to perform the analysis.

3.2.1 Advantages

- Time saving
- Accurate results

4. ALGORITHMS

Having the sales data of the retail store, the proposed work suggests the following various steps for predicting the sales of different categories available. The architectural diagram for the proposed algorithm is shown in Figure 1. The various steps involved are explained hereunder.

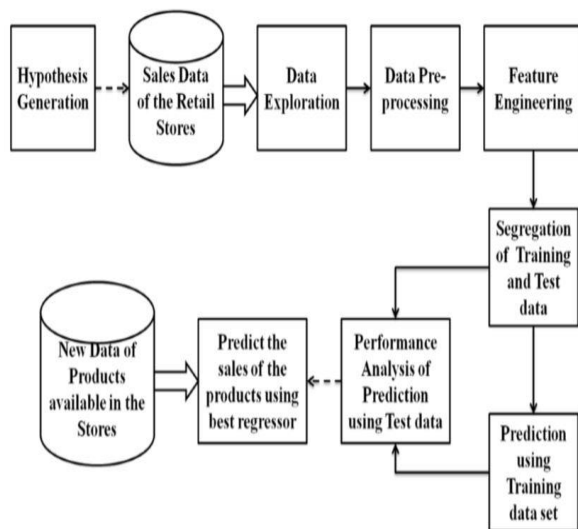


Fig 4.1: Architectural Diagram

When we consider a business problem, we try

to achieve more accuracy by changing and implementing different models. But, after a certain point, we notice that we will be struggling to improve the accuracy of the model. To overcome such type of problems data exploration comes into the picture. The first step in Data exploration is to look into the dataset and to discover the information regarding the available and the hypothesized data. The dataset under consideration has the following features as the variables as shown in Table 1. We can see that there are 6 features which are hypothesized and present in the dataset, 3 features present in the dataset but not hypothesized and 9 features hypothesized and not found in the data. This can be best depicted in the diagram as shown in below

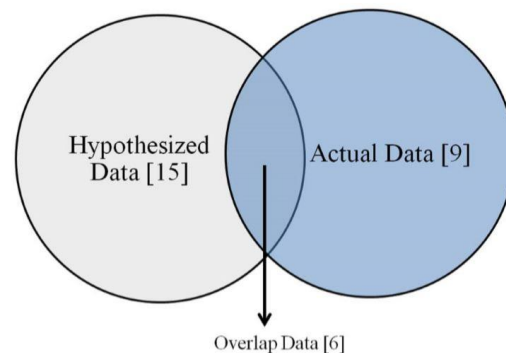


Fig 4.2 : Venn Diagram

The dataset under consideration has some values missing in the columns “Outlet_Size” and “Item_Weight”. The missing values of the data will be imputed in the data preprocessing section. The variables which are present in our dataset can be grouped as categorical variables and numerical variables.

Variable	Description
Item_Identifier	Product ID
Item_Weight	Weight of Product
Item_Fat_Content	Fat content in Product
Item_Visibility	% of the display area occupied
Item_Type	Category of the product
Item_MRP	Cost of the product
Outlet_Identifier	Store ID
Outlet_Establishment_Year	Year of establishment
Outlet_Size	Ground area covered by the store
Outlet_Location_Type	Type of city store located
Outlet_Type	Type of the store
Item_Outlet_Sales	Sales of product in a particular store

Table 6.1 : Data Description

4.1 Multiple Regression :

Multiple Regression is an extension of simple linear regression. It is used when we want to predict the value of the dependent variable based on the value of two or more independent variables. In general, multiple linear regression procedures will estimate the value of the variable based on the following equation –

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n$$

Here, a represents the coefficients.

4.2 Polynomial Regression :

Polynomial Regression is the form of regression analysis, in which the relationship between the independent variable and the dependent variable is modeled as an nth degree polynomial. This is a non-linear regression model and polynomial regression is the extension of simple linear regression with the order being equal to 1. The polynomial regression procedures will estimate the value of the variable based on the following equation –

$$y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

Here, a represents the coefficients.

4.3 LASSO Regression :

LASSO regression in machine learning stands for Least Absolute Shrinkage and Selection Operator. It performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability. The major objective of LASSO regression is to solve

$$\min_{\beta_0, \beta} \left\{ \frac{1}{N} (y - \beta_0 - x_i^T \beta)^2 \right\}$$

Where β is subjected to the following constraint and t is the free parameter which determines the amount of regularization.

$$\sum_{j=1}^p |\beta_j| \leq t$$

4.4 Ridge Regression :

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation since we must somehow indicate whether the variables in a particular formula are standardized or not. As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is on a standardized scale.

$$Y = XB + e$$

Where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors or residuals.

4.5 AdaBoost :

AdaBoost stands for Adaptive Boosting. This algorithm is mainly used in improving the performance of the model created. In this algorithm the final output is obtained by taking the weighted sum of the outputs of the weak learners. In this algorithm the output is

manipulated in such a way that, the algorithm favors the instances which were mispredicted by the weak learners. Hence making the algorithm adaptive. AdaBoost refers to a particular method of training a boosted model. Where the boosted model is in the form –

$$F_T(x) =$$

$$\sum_{t=1}^T f_t(x)$$

In the above formula f_t represents a weak learner and x represents an object which is as an input to the weak learner. The weak learner f_t returns the predicted value of the object. $h(x_i)$ is the hypothesis generated by each weak learner for every data sample in the training set. A coefficient α_t is assigned for every weak learner $h(x_i)$, which is selected at iteration t such that the sum of training error E_t is minimized.

$$E_t =$$

$$\sum_i E|F_{t-1}(x_i) + \alpha_t h(x_i)|$$

5. Results:

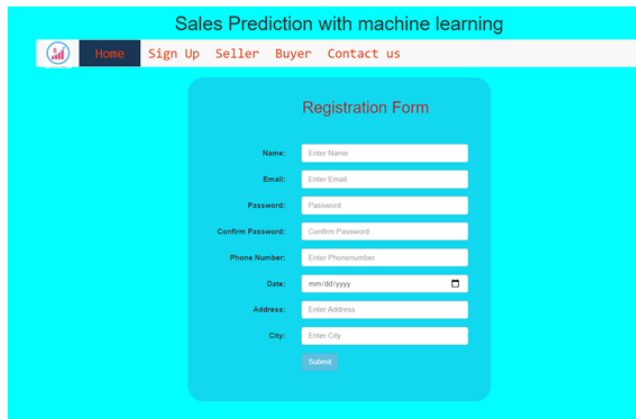


Fig:5.1 SIGN UP PAGE

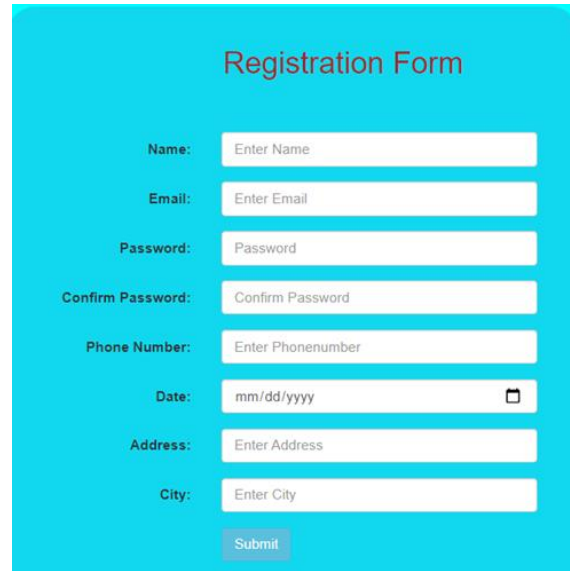


Fig 5.2 REGISTRATION PAGE

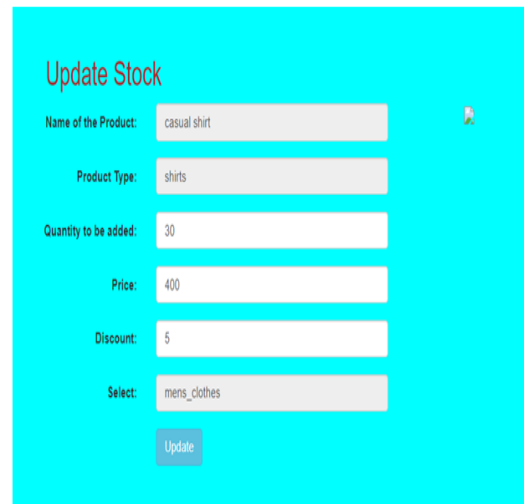


Fig 5.3 UPDATE STOCK



Fig5.4 PREDICTION BAR G

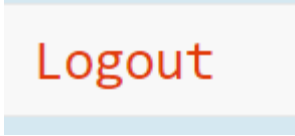


FIG5.7 LOG OUT

Fig:5.5 BUY PRODUCT

Ordered Products

Product ID	Product Name	Product Price	Discount	Order Price	Quantity	Order Date	Category
PTC041078	400	shirts	5	390	1	Tue Jul 14 15:53:2021	mens_clothes

Fig: 5.6 ORDERED PRODUCT

Fig CONTACT US



6. CONCLUSION

In this project, we compare the performance of different algorithms on store sales dataset and analyze the algorithm with the best performance. Here we have observed that the AdaBoost algorithm has the highest RMSE value of 1350.72 and the algorithm with the least RMSE value is GradientBoost having 1088.64. The algorithm in terms of highest R^2 is GradientBoost with the value of 0.59 and the algorithm with the least R^2 is AdaBoost with the value of 0.40. Hence, by the obtained results we can see that the GradientBoost algorithm is the best predictor for the considered dataset having the least RMSE value of 1088.64 and the highest R^2 value of 0.59. We can also conclude that without proper hyper parameter tuning the AdaBoost algorithm won't be able to perform as expected and the performance deteriorates

REFERENCES

- [1] Ching Wu Chu and Guoqiang Peter Zhang, "A comparative study of linear and nonlinear models for aggregate retails sales forecasting", *Int. Journal Production Economics*, vol. 86, pp. 217-231, 2003.
- [2] Giuseppe Nunnari, Valeria Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study", *Proc. of IEEE Conf. on Business Informatics (CBI)*, July 2017.
- [3] <https://halobi.com/blog/sales-forecasting-five-uses/>. [Accessed: Oct. 3, 2018]
- [4] Zone-Ching Lin, Wen-Jang Wu, "Multiple LinearRegression Analysis of the Overlay Accuracy Model Zone", *IEEE Trans. on Semiconductor Manufacturing*, vol. 12, no. 2, pp. 229 – 237, May 1999.
- [5] O. Ajao Isaac, A. AbdullahiAdedeji, I. Raji Ismail, "Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis", *Int. Journal on Mathematical Theory and Modeling*, vol. 2, no. 2, pp. 14 – 23, 2012.
- [6] C. Saunders, A. Gammernan and V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables", *Proc. of Int. Conf. on Machine Learning*, pp. 515 – 521, July 1998.*IEEE TRANSACTIONS ON INFORMATION THEORY*, VOL. 56, NO. 7, JULY 2010 3561.
- [7] "Robust Regression and Lasso". HuanXu, Constantine Caramanis, Member, IEEE, and ShieMannor, Senior Member, IEEE. 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration."An improved Adaboost algorithm based on uncertain functions".ShuXinqing School of Automation Wuhan University of Technology.Wuhan, China Wang Pan School of the Automation Wuhan University of Technology Wuhan, China.
- [8] XinqingShu, Pan Wang, "An Improved Adaboost Algorithm based on Uncertain Functions", *Proc. of Int. Conf. on Industrial Informatics – Computing Technology, Intelligent Technology, Industrial Information Integration*, Dec. 2015.
- [9] A. S. Weigend and N. A. Gershenfeld, "Time series prediction: Forecasting the future and understanding the past", Addison-Wesley, 1994.
- [10] N. S. Arunraj, D. Ahrens, A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting, *Int. J. Production Economics* 170 (2015) 321-335.
- [11] D. Fantazzini, Z. Toktamysova, Forecasting German car sales using Google data and multivariate models, *Int. J. Production Economics* 170 (2015) 97-135.
- [12] X. Yua, Z. Qi, Y. Zhao, Support Vector Regression for Newspaper/Magazine Sales Forecasting, *Procedia Computer Science* 17 (



2013) 1055–1062.

[13] E. Hadavandi, H. Shavandi, A. Ghanbari, An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: a Case study of the printed circuit board, *Expert Systems with Applications* 38 (2011) 9392–9399.

[14] P. A. Castillo, A. Mora, H. Faris, J.J. Merelo, P. GarciaSanchez, A.J. Fernandez-Ares, P. De las Cuevas, M.I. Garcia-Arenas, Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment, *Knowledge-Based Systems* 115 (2017) 133-151.

[15] R. Majhi, G. Panda and G. Sahoo, “Development and performance evaluation of FLANN based model for forecasting of stock markets”.*Expert Systems with Applications*, vol. 36, issue 3, part 2, pp. 6800-6808, April 2009.