Store Sales Prediction Using Machine Learning and AI

Shinde Saurabh,
Karanjkar Vaibhav,
Shirsath Raj,
Vibhute Vaishnavi

Research Scholar, Department of Computer Engineering, Sapkal College of Engineering Nashik, Maharashtra, India
Research Scholar, Department of Computer Engineering, Sapkal College of Engineering Nashik, Maharashtra, India
Research Scholar, Department of Computer Engineering, Sapkal College of Engineering Nashik, Maharashtra, India
Research Scholar, Department of Computer Engineering, Sapkal College of Engineering Nashik, Maharashtra, India

ABSTRACT

Sales forecasting is one of the main issues of supply chains. It aimed to optimize stocks, reduce costs, and increase sales, profit, and customer loyalty. For this purpose, historical data can be analyzed to improve demand forecasting using various methods like machine learning techniques, time series analysis, deep learning models, artificial neural networks, etc. In this work, an intelligent demand forecasting system is developed. This improved model is based on analyzing and interpreting the historical data by using different forecasting methods, including time series analysis techniques, support vector regression algorithms, and deep learning models. The organization sells gifts primarily on the online platform. The customers who make a purchase consume directly for themselves. Small businesses buy in bulk and sell to other customers through the retail outlet channel. The data contains information about more than 1000 stores in the country and there are more than 1000000 entries in the data. We are required to predict the sales for the Store-Day level for one month

Keyword: - *Correlation, Machine Learning, Performance Metrics, Sales Forecasting*

1. INTRODUCTION

Sales forecasting can be defined as the prediction of upcoming sales based on the past sales that occurred. Sales forecasting is of paramount importance for retail stores which are entering new markets or are adding new services, products, or which are experiencing high growth. The main reason a retail store does a forecast is to balance marketing resources and sales against supply capacity planning. We are using this concept of sales forecasting to forecast or predict the sales of a retail store on monthly basis. 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 retailers allocate significant financial and human resources to perform this task genuinely, which requires a 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 predictions by the model. Retailers 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 retail’s services or products. Forecasting can be used to predict sales revenue at a product level, or an individual business level, or a company level. In this project, we have concentrated on sales predictions based on historical data of the stores. Future sales plan aids in optimal utilization of the facility, scheduling, conveyance, and effective control of inventory. These, in turn, result in the enhancement of customers satisfaction. In the recent past, many investigations

addressing the problem of sales forecasting have been reported. Sales forecasts affect a retail's marketing plan directly. The marketing department is responsible for how retailers and their customers interpret 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 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 developing productive models 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 algorithms like Ada-Boost, Gradient Tree Boosting to get the maximum accuracy. Multiple Regression is a statistical tool used to predict the output which is dependent on several other independent predictors or variables. 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 nonlinear relationship between the independent variable x and the dependent variable y. Here the model usually fits the variable y as the nth degree of variable x. Ridge Regression is a way to create a model when the predictor variable has multi-co-linearity. Lasso is a regression method that performs both regularization and variable selection to enhance prediction accuracy. The Ada-Boost is a boosting algorithm that 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 arrives at the output. Ada-Boost can significantly improve learning accuracy no matter whether applied to manual data or real data. The elastic net method overcomes the limitations of the Lasso. In this project, a wide range of forecasting methods is implemented because the combination of multiple forecasts can be used to increase forecast accuracy.

2. PREVIOUS WORK

A lot of work has been done related to sales prediction, as it's one of the most important concern by the retailer. Sales prediction could be done by customer related features, store related features, and item related feature. For example, (Chen, Lee, Kuo, Chen, & Chen, 2010) has been working on forecasting sales model on fresh food. The prediction is based on both item and customer, as when consumers are making purchases of food products, they would first consider if the foods are fresh and if they are expired. But the methodology can't feed into our problem, as only store related features are provided and customer-item prediction could not be predicted at all. Another retail sale prediction problem has been described in (Giering, 2008). It's also a sales prediction problem based on customer related feature and item related feature where SVD and recommendation system is applied. Although the methodology can't be well applied in our problem, there is still inspiration from their work: using log(Sales) as the target in prediction as it might normalized the distribution.

(Chang, Liu, & Lai, 2008) has described a way to make sales prediction only based on item related feature which is more similar to our problem where only store related feature is provided. In (Chang et al., 2008), it obtained *case-based reasoning model* and *k-nearest neighbours' algorithm* to find the most similar item with sale history, given an item without sale history. We tried *k-nearest neighbours' algorithm* in our dataset to find the most similar store and time information. We got some result; however, its performance is not as good as expected. The model and the result could be found in the following section.

(Thiesing, Middelberg, & Vornberger, 1995) has adapted *Back-Propagation* as a neural network method to make sales prediction on Transputer system. The article has also described how they applied parallel computing into the model to improve the efficiency of computing.

3. DATASET INFORMATION

Rossman operates over 3,000 drug stores in 7 European countries. The task is to predict 6 weeks of daily sales for 1,115 stores located across Germany. 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 motivation.

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
0	1	5	2015-07-31	5263	555	1	1	0	1
1	2	5	2015-07-31	6064	625	1	1	0	1
2	3	5	2015-07-31	8314	821	1	1	0	1
3	4	5	2015-07-31	13995	1498	1	1	0	1
4	5	5	2015-07-31	4822	559	1	1	0	1

Fig -1: Dataset Snapshot

Store ID
It is customary to think the store ID as one feature because sales may change from store to store.

Day of Week
It's also easy to think that in different day of week, every store will have different sales since people get used to shop in different days.

Numbers of Customers
In the training dataset which represents the past information includes the numbers of customers. However, in the test dataset which provides the future information doesn't have the numbers of customers.

Open
"Open" shows whether this store is open or not in a specified day.

Promo
It indicates whether a store is running a promo on that day.

School Holiday
It shows the information whether there was holiday for the school on that day.

State Holiday
It shows the information whether there was state holiday on that day.

Table -1: Dataset Statistic

STATISTICS	NUMBERS
Dataset size	1017209
Testing data size	41088
Total stores number	1115
Training data Time ranges	2013-01-01 to 2015-07-31
Testing data Time ranges	2015-07-31 to 2015-08-31

4. EXPLORATORY DATA ANALYSIS

Here are some insights that could be found in our data.

4.1 Top-10 Store sales

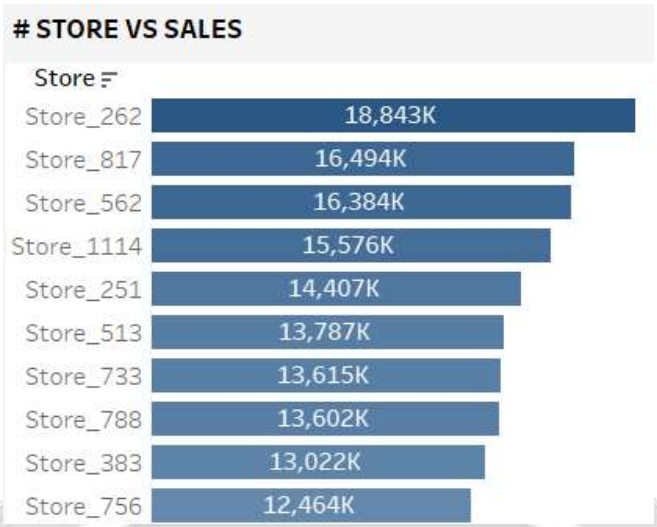

Fig -2: Top 10-Store Sales

From the dataset we had extracted top 10 sales who had made the maximum sales in 2 and half year of span. From fig-2, we can observe that store with store id Store_262 had made highest sales in 2 and half year of span.

4.2 Monthly Sales of Stores

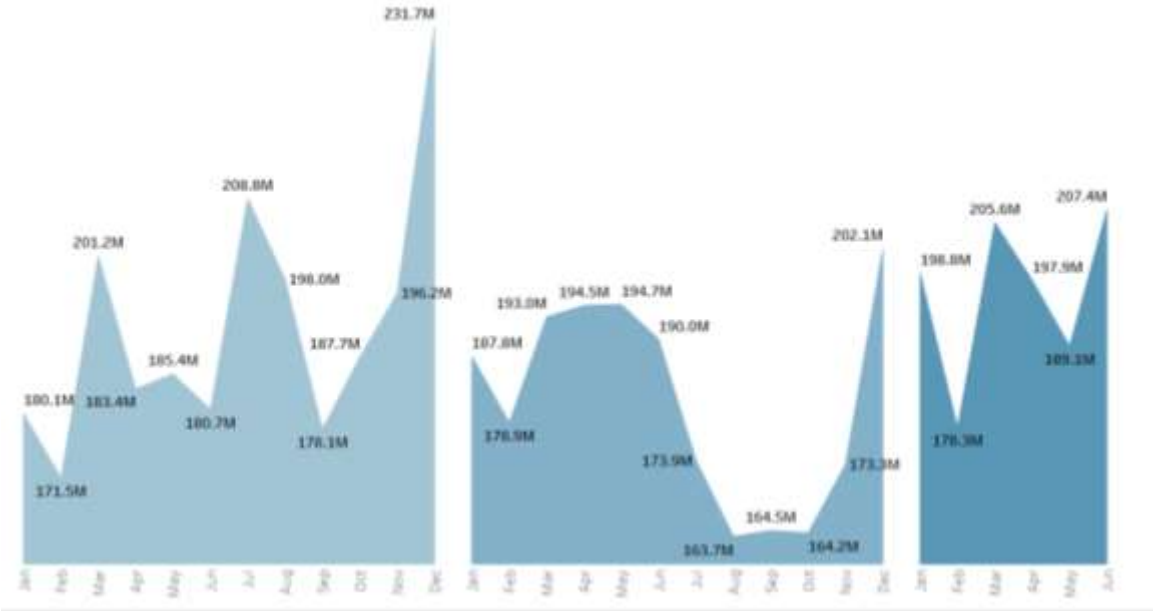

Fig -3: Monthly Sales Analysis

Total sales of store from 2013 to 2015 have been visualized in fig -3. We can say that, in the month Dec 2013, sales are maximum. While in the span of Aug 2014 and Sep 2014 Sale was minimum.

4.3 Day wise Customer and sales

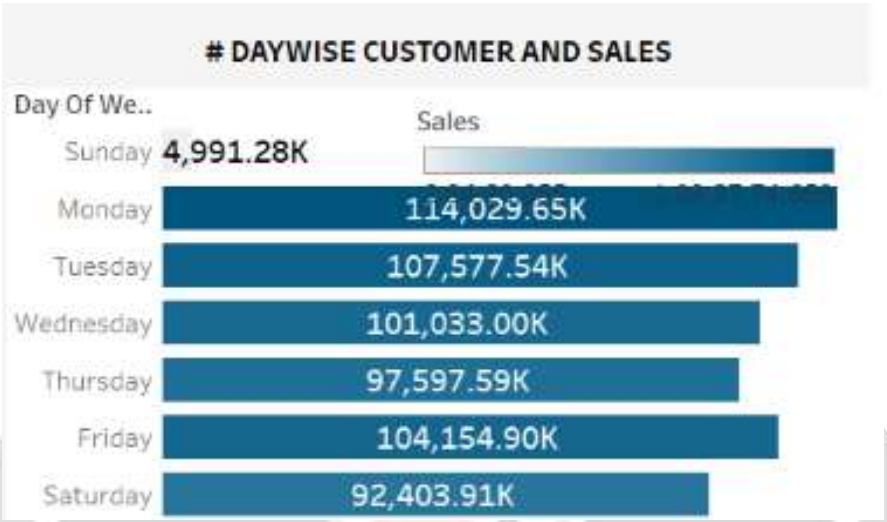

Fig -4: Day wise Customers and sales

In fig, it is observed that, sale was minimum on Sunday and maximum on Monday. Maybe it's because most of the stores stay closed on Sunday.

5. METHODOLOGY

5.1 Linear Regression

Simple linear regression is useful for defining a relationship between two continuous variables. One is an indicator or independent variable and another is an answer or dependent variable. It looks for a statistical relationship, but not a deterministic one. The relationship between the two variables is said to be deterministic if one variable can be precisely represented by the other. For example, it is possible to correctly forecast Fahrenheit by using temperature in degree Celsius. The mathematical equation is not sufficient to assess the association between the two variables. For example, the relationship between weight and height. The Equation for the Simple linear regression is: $Y = a + bX$ where Y is the expected value of the dependent variable y for every specified value of the independent variable X, a is the intercept, b is the regression coefficient and X is independent variable

5.2 Gradient boosting Regression

Gradient boosting is some kind of enhancement in Machine Learning. It is based on the premise that, when combined with previous ones, the best possible current iteration will minimize the maximum prediction error. The key idea for this next iteration is to set the target outcomes to minimize the error. One of the most successful Machine Learning models for predictive analytics is the Gradient Boosted Regression Trees (GBRT) model (also called Gradient Boosted Machine or GBM), which makes it an industrial workhorse for Machine Learning. The Boosted Trees Model is a type of additive model that combines decisions from a sequence of base models to make predictions. One can write this class of models more formally, as: $g(x) = f0(x) + f1(x) + f2(x) + f3(x) + ...$ Where the final classifier g is the amount of the specific classifiers. Each base classifier is a simple decision tree for model boosted trees. This broad approach of using multiple models is called model ensemble to achieve better predictive

performance. Unlike Random Forest, which independently builds all the base classifiers, each using a subsample of data, GBRT uses a particular technique of assembly called gradient boosting.

5.3 Ridge regression

Ridge regression uses the l2 regularization which allows to create a model when the number of predictor variables in a set exceeds the no. of observations. It is able to work with multi collinear data. It does not face the problem of over fitting. Here the penalty is on the sum of squared coefficients

5.4 Random Forest Regression

Random Forest is one of the most powerful Machine Learning frameworks for predictive analytics. A random forest method is a type of discrete structure that allows predictions by integrating decisions from a series of simple models. More formally, this subset of models can be written as: $g(x) = f0(x) + f1(x) + f2(x) + f3(x) + ...$ Where the initial configuration g is the number of the initial specific model's fi. Here, any base classifier is a simple decision tree. This wide-ranging technique of using multiple models to improve predictive performance is called model assembling. In random woods, all baseline models are built independently using a separate subset of results.

5.5 XGBoost

XGBoost is an implementation of Gradient Boosted decision trees. This library was written in C++. It is a type of Software library that was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning. XGBoost models majorly dominate in many Kaggle Competitions.

In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. Weight of variables predicted wrong by the tree is increased and these the variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

5.6 Principal component analysis

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D. PCA helps in finding a sequence of linear combinations of variables.

In the above figure, we have several points plotted on a 2-D plane. There are two principal components. PC1 is the primary principal component that explains the maximum variance in the data. PC2 is another principal component that is orthogonal to PC1.

5.7 Time Series Modelling

An ordered set of observations with respect to time periods is a time series. In simple words, a sequential organization of data accordingly to their time of occurrence is termed as time series.

For example, "how do people get to know that the price of an object as increased or decreased over time", they do so by comparing the price of an object over a set of the time period.

A time series data is the set of measurements taking place in a constant interval of time, here time acts as independent variable and the objective (to study changes in a characteristics) is dependent variables.

5.8 Artificial neural networks

The artificial neural network is a computer networking system that can perform huge and intelligent tasks. It is a parallel and distributed processing system that can accomplish most complex tasks of recognition, prediction and detection without increasing the complexity of the problem. The artificial neural network has one input layer and one output layer between which are the hidden layers that process data. Each layer by processing the data forwards the result to the next hidden layer and Nally the output layer obtains the result after the data processing. The artificial neural network is one the most popular machines of artificial intelligence that are used almost in every _eld nowadays. It uses certain different models for processing data like feedforward back propagation, NARX model with different functions for each model. These are dynamic machines capable of solving complex to everyday problems and made the human life easy

6. RESULTS AND DISCUSSION

We have divided a problem into two parts. In the first part we are not converting columns into dummy variables. You can see the result of that below.

Table -2: Results of data without stores dummies variables

	Training			Testing		
Algorithm	**RMSE**	**MAE**	**R-Squared**	**RMSE**	**MAE**	**R-Squared**
Linear Regression	2621.13	1776.86	0.14	2549.40	1804.18	0.02
Decision Tree	2420.80	1599.30	0.35	2558.5	1780.91	0.19
Random Forest	2406.04	1591.96	0.35	2551.66	1776.22	0.22
Ridge	2621.13	1776.86	0.14	2549.39	1804.19	0.02
Gradient Boosting				1603.34	1143.96	0.71

Table-3: Results of Artificial Neural Network Model

Model	**No. of Neurons**	**Dropout**	**Optimizer**	**RMSE**	**MSE**
ANN model 1	128 + 64 = 192	0.2 + 0.2	Adam	2572.72	1825.57
ANN model 2	128 + 64 + 642 = 256	0 + 0.2 + 0.4	Adam	2540.78	1665.93
ANN model 3	100 + 64 = 164	0 + 0.2	Adam	3144.27	2097.51

It could be seen in table-2 that Gradient boosting algorithm gives the best performance among all of them with validation RMSE 1603 and MAE 1143.

In the second case, in which we have created dummy columns for each categorical values in store. In this case, dimension of the dataset has increased to 1124 columns in total. Due to the insufficient ram, it was throwing memory error. That's why we drop all the entries of 2013 from the dataset. Performance was compared with the model which was having all entries. Change in accuracy was negligible, so we kept data in the same way and applied various algorithms on the same.

Table -3: Results of data with stores dummies variables

	Training	**Testing**

Algorithm	**RMSE**	**MAE**	**R-Squared**	**RMSE**	**MAE**	**R-Squared**
Linear Regression	1503.49	1077.31	0.69	1213.39	860.05	0.88
Ridge	1503.49	1077.31	0.69	1213.14	860.04	0.88
Decision Tree	2545.79	1898	-1.06	2307.95	1609.55	0.30
Random Forest	2902.90	2171.49	-87.10	2572.09	1804.92	-0.18
Random Forest After PCA	2631.79	1921.01	-2.52	2427.02	1668.89	0.14
XGBoost1	1548.05	1112.51	0.67	1251.69	884.067	0.87
XGBoost2	1502.92	1164.87	0.80	1502.92	1164.87	0.80

It could be seen that Linear Regression and Ridge regression algorithm gives the best performance among all of them with validation RMSE 1213 and MAE 860.

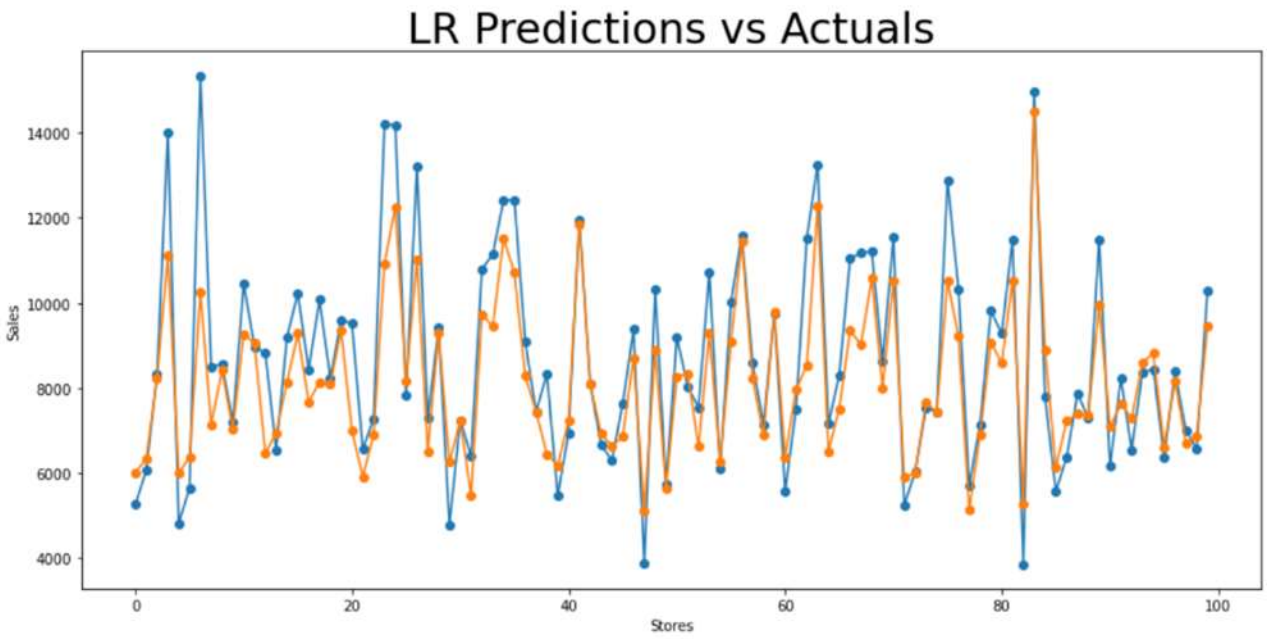

Fig -5: Predicted vs actual comparison of Linear regression model

7. CONCLUSIONS

From the result it can be conclude that accuracy exponentially increases after creation of dummy variables of store column. In the first case, linear regression model was predicting the test results with RMSE score of 2549.40, and MAE of 1804.18. But after creating dummy columns, data becomes more informative and meaningful for the

algorithm. Results exponentially changes to RMSE score 1213.39 and MAE 860.05 which is the best accuracy given for our problems. We have also implemented some deep learning concepts, but results were not up to the mark. So finally, we can say that store sales can be predicted by using machine learning algorithms with a good accuracy.

8. 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 Linear Regression 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. 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.

[6] C. Saunders, A. Gammerman 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] A. Koutsoukas, K. J. Monaghan, X. Li and J. Huan, Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modelling bioactivity data," Journal of Cheminformatics, 2017. [8] Xinqing Shu, 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.