Investigations on E-commerce Data for Forecasting the Efficient Promotional Platform Using Supervised Machine Learning

Rajeev Kamal
Innovation lab
Dayananda Sagar University
Bengaluru, India
rajeev-ece@dsu.edu.in

Abhinav Karan
Innovation lab
Dayananda Sagar University
Bengaluru, India
abhinav-cse@dsu.edu.in

Arungalai Vendan S
Innovation lab
Dayananda Sagar University
Bengaluru, India
arungalai-ece@dsu.edu.in

***Abstract*— The technological advancements provide various platforms for e-commerce companies to leverage the maximum benefits for higher revenue generation. Websites and product applications are the prominently used platforms by the users or customers to identify the various features of the product while also adopting this mode for purchase. E-commerce industries consider the determination of the most ideal platform as a critical task. Datas pertaining to user's inclination towards website or product applications are collected through registration credentials. With the available data which is big in size, it becomes a herculean task to precisely predict. Nevertheless conventional tools have been adopted by the industries whose results are not satisfactory and promote further explorations with advanced tools. Accounting for these complex tasks, this study attempts to employ linear regression, a versatile tool for predicting the most popular platform that generates maximum revenue. Various machine learning libraries have been used to scrutinize each and every parameter like number of registrants and the frequency of visits to arrive at a predicted value that would facilitate e-commerce industries for promoting the specific business platforms and eventually reap higher revenue. The proposed technique may be extrapolated for parametric forecasting by any industry with local customization.**

Keywords— Linear Regression, Product Application, Scikit, Matplot, Tensorflow

I. INTRODUCTION

Linear regression is statistical analysis technique that attempts to create a predictive model by establishing the relationship between dependent and independent variable in the form of a straight line. The line tends to project trends in data on various applications such as prices of different entities, detecting different kind of diseases and many more. There are various ways to calculate linear regression, one of the popular method is least square method. Traditional methods calculations to perform linear regression can be quite complex. With the advent of various calculation packages, the time taken to implement linear regression has become relatively low. Researchers have used linear regression for several applications and few of those have been presented in brevity in this section. Hirose et al (2012) [1] proposed an efficient technique to forecast accurate results by combining k-NN regression and linear regression methods. The experiment conducted in this paper by applying the proposed techniques resulted in an efficient and improved estimation of prices for auction of used cars.Wang et al (2017) [2] applied least square method to achieve the regression equation. The error reproduced by the model is minimized by applying methods such as total sum of squares, residual sum of squares, regression sum and model error. Kavitha, S. et al (2016) [3] proposed techniques to forecast the growth of business in the future by analyzing the parameters such as consumers interest, behaviour and product profit. The data referred in this research work is of nature time series and linear regression is considered to one of the best fit for giving accurate prediction on time series data. Bayindir et al (2011) [4] proposed a low cost and simple solution for power factor correction by developing a technique with the application of both linear regression and ridge regression method. The technique was evaluated on the basis of 10-fold Cross Validation protocol, which suggested that linear regression is more suitable as compared to the ridge regression based on the performance results. Men et al (2009) [5] proposed a multiple linear regression technique to determine Pb2+ and Cd2+ applied upon ion selective electrodes. Regression equation were determined separately for both the metals by considering regression coefficient to be 0.9999 which yielded acceptable results. Guo et al (2009) [6] developed an improved version of multiple linear progression to forecast oilfield output. The suggested model discussed in the paper was created by applying rigorous statistical calculation and experiments. The forecasted results showed good coherence with the measured output. Harimurti, R. et al (2018) [7] projected a method to forecast students psychomotor domain. Based on the experiments conducted using various regularization techniques and subsequent measurement of performance using cross-validation and random sampling, it was concluded that elastic net regression technique is more appropriate to forecast errors with minimal errors. Peng et al (2018) [8] proposed a technique, which attempts to estimate the risk premium on the assets by applying multiple linear regression. Multiple macroeconomic factors were considered while performing analysis to estimate the risk premium. Experiments revealed that the generalized approach of moment estimators of risk premiums yield appreciable results on individual assets over historical averages. Feng et al (2017) [9] proposed a technique to predict the bike rental program under capital bike share program in Washington, D.C. The random forest model suggested by the author and a GBM packed enhanced the efficacy of the decision tree. The accuracy of the accomplished results had substantially increased while improving the reliability of bicycle rental prediction. Naseem, A. et al (2009) [10] implemented linear regression for face

978-1-7281-0630-4/19/$31.00 ©2019 IEEE

identification. Standard databases were used to evaluate the algorithm and comparative study was performed by applying benchmarking algorithm to justify the efficiency of the proposed technique.

The literature survey highlights the features and applicability of linear regression for versatile applications. However, implementations using sophisticated libraries such as tensorflow, keras, pytorch etc generally incorporated in machine learning have been inadequately used. Nevertheless conventional tools have been adopted by the industries whose results are not satisfactory and promote further explorations with advanced tools. Considering the benefits and the wider option provided by these libraries, this investigation attempts to apply scikit-learn, pandas, matplotlib and seaborn libraries to develop platform to predict the cost and efficiency functions of e-commerce industry individually for websites and product application options presented by them.

II. Data Sets and Work-flow

The focus of this paper is to propose an efficient prediction method for an Ecommerce company on putting their efforts on mobile app or website in order to improve the shopping experience for their customers.

The proposed method for prediction is implemented using Linear Regression by applying Scikit-learn library on a real time dataset from a ecommerce company to forecast user experience. Linear regression model is implemented using Jupyter Notebook along with some data analysis and data visualization library. Firstly, the data set is imported which gives information about the customer info such as Email, Address, and their color Avatar. The data set comprises of some additional columns like Avg. Session Length, Time on App, Time on Website, and Length of Membership. The additional parameters mentioned play a vital role in forecasting the customer experience. The information from the dataset about the rows and columns is extracted by info ( ) method on customers data frame. Table 1 mentioned below shows the result after applying info method on customers data frame.

TABLE I. Customers Data Frame

Features of Data-set	***Number and Types***
E-mail	5000 non-null float 64
Address	5000 non-null float 64
Avatar	5000 non-null float 64
Avg. Session Length	5000 non-null float 64
Time on App	5000 non-null float 64
Time on Website	5000 non-null float 64
Length of Membership	5000 non-null float 64
Yearly Amount Spent	5000 non-null object

Libraries such as seaborn and matplotlib are applied on the available data frame to visualize the data. In the proposed work initial visualization is done as shown in figure 1 to generate a joint plot based on the two parameters namely Time on Website and Yearly Amount Spent. Thorough interpretation of the scatter plot is unavailable. Since the scatter plot did not. clearly indicate any predictions, another joint plot is generated as shown in figure 2 by modifying the x axis parameter as Time on app instead of Time on Website. Figure 2 results shows some minor improvements in visualization as it creates better correlation between the two parameters. Further to the previous plots, another improved version of joint plot has been applied to generate 2D hex bin to show comparison between Time on App and Length of Membership as shown in figure 3

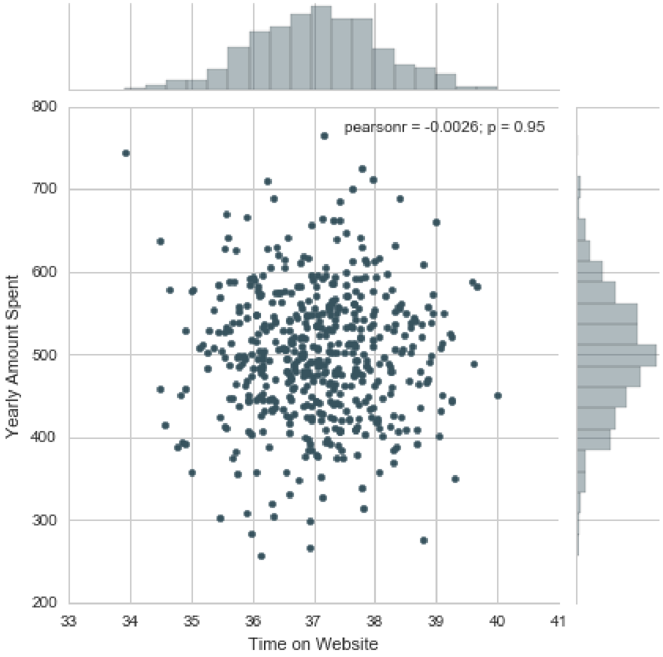

Fig. 1. Joint Plot - Time on Website v/s Yearly Amount Spent.

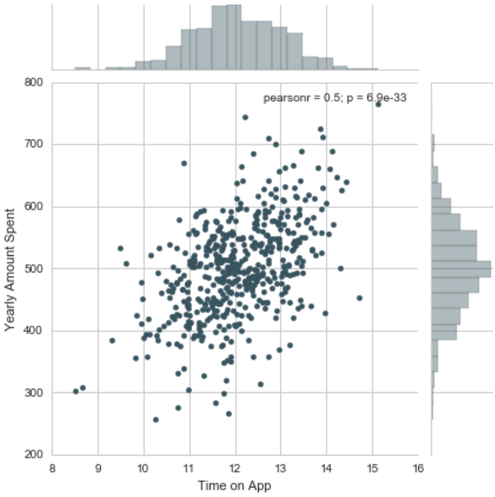

Fig. 2. Joint Plot - Time on App v/s Yearly Amount Spent

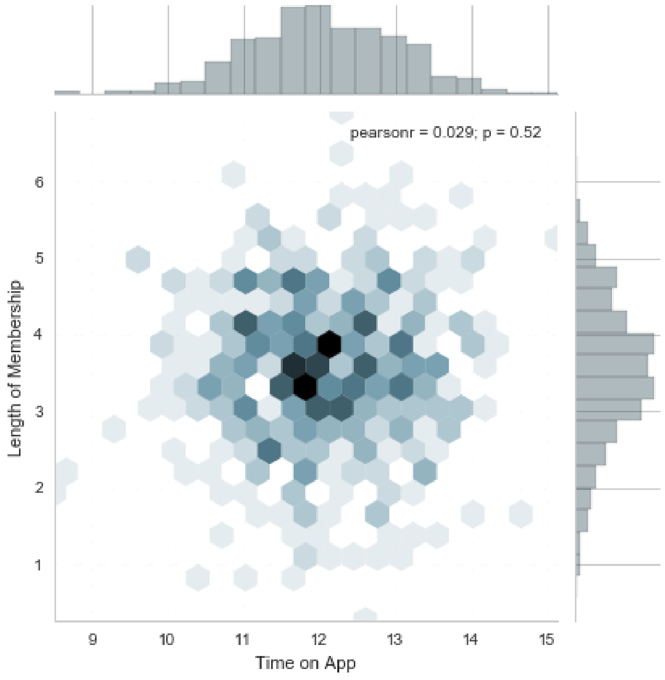

Fig. 3. Joint Plot 2D Hex Bin - Time on App v/s Length of Membership

Subsequently, in order to explore all the correlations among all the parameters present in the available data frame, a pair plot is generated as shown in figure 4.

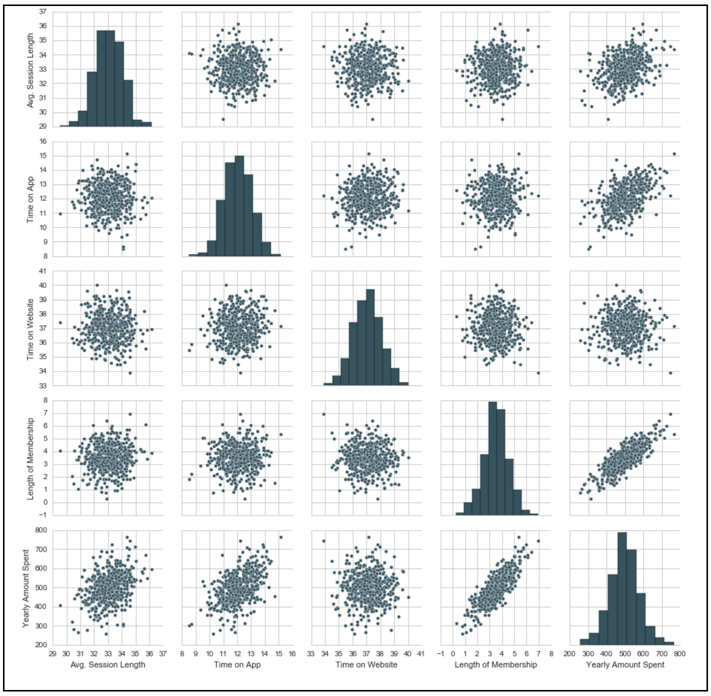

Fig. 4. Pair Plot

From figure 4, it may be inferred that can be made is yearly amount spent is proportional with Length of Membership. After establishing the correlation, a linear model has to be plotted using seaborn library. The plot shown in figure 5 estimates the best linear fit between the two parameters with minimum error. The plot illustrates that the tenure of membership is proportional to the annual expenditure.

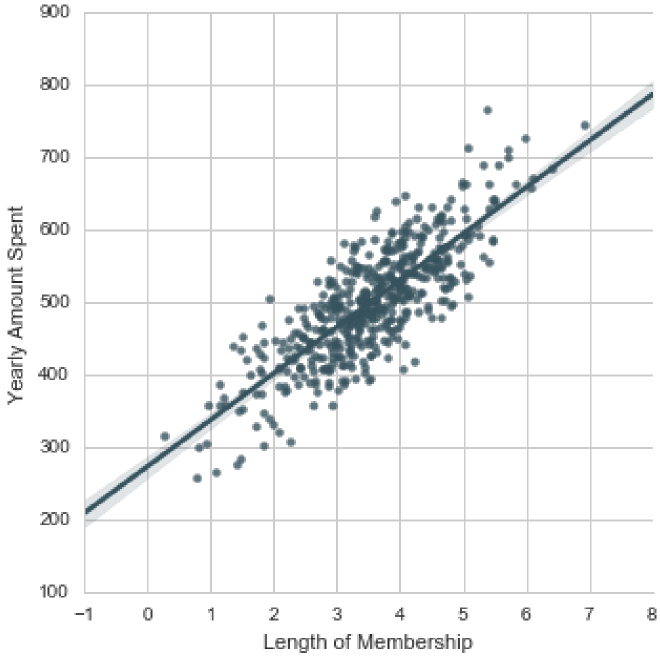

Fig. 5. Linear Plot - Length of Membership v/s Yearly Amount Spent

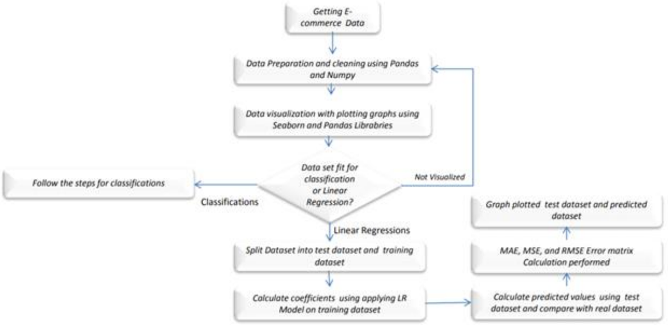
Getting E-commerce Data
Data Preparation and cleaning using Pandas and Numpy
Data visualization with plotting graphs using Seaborn and Pandas Librabries
Data set fit for classification or Linear Regression?
Not Visualized
Follow the steps for classifications
Classifications
Linear Regressions
Split Dataset into test dataset and training dataset
Calculate coefficients using applying LR Model on training dataset
Calculate predicted values using test dataset and compare with real dataset
MAE, MSE, and RMSE Error matrix Calculation performed
Graph plotted test dataset and predicted dataset

Fig. 6. Flowchart for the Work flow

Based on the previous analysis two features are shortlisted, which are the best fit to develop linear regression models by applying Sci-kit learn library. The first step towards developing the machine learning model is to divide the entire data set into train and test data. Before splitting some prerequisites are to be established such as initialization of some variable to some parameters of the dataset.

train test split( ) module from Sci-Kit learn library splits the available dataset by allocating 30 percent to test data and the remaining to training data. There is a possibility of obtaining different data subset of train and test on every execution of splitting, which are be avoided by providing a fixed value to variable random state.

All these steps are sequentially illustrated in the form of flowchart shown in figure 6, which explained us on the process of developing a trained model using linear regression to forecast accurate data based on the predictions.

III. Result and Analysis

After proper splitation of datset, the model is trained. Further to training, evaluation of the model trained are performed by calculating the coefficients of the features in the model as shown in table 2.

TABLE II. Coefficients of features

Coefficients
25:98154972
38:59015875
0:19040528
61:27909654

The evaluation data frame shown in table 3 illustrates the coefficient value with respect to the features used for predictions.

TABLE III. Evaluation of data Frame

Features	Coefficient
Avg. Session Length	25:98154972
Time on App	38:59015875
Time on Website	0:19040528
Length of Membership	61:27909654

With respect to the data frame shown in table 3, the coefficient values are interpreted as follows :

- A unit increase in Avg. Session Length resulted in an overall increase of 25.98 dollars spent.
- A unit increase in Time on App resulted in an overall increase of 38.59 dollars spent.
- A unit increase in Time on Website resulted in an overall increase of 0.19 dollars spent.
- A unit increase in Length of Membership resulted in an overall increase of 61.27 dollars spent.

After evaluating the model based on the coefficients generated, experiments are conducted by passing test data to the trained model to check to check the forecasting accuracy. The predictions are illustrated from the generated scatter plot as shown in the figure 7.

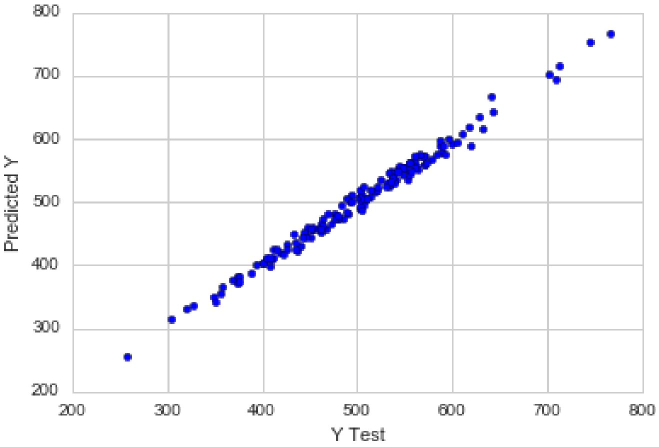

Fig. 7. Scatterplot: Prediction

To analyze the forecasting a histogram has been plotted as shown in figure 8.

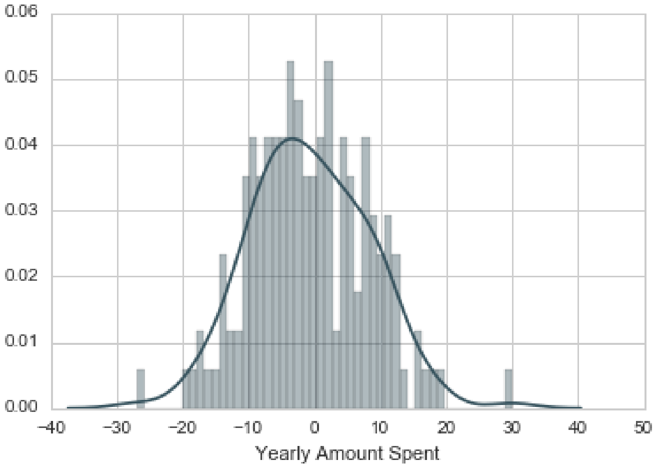

Fig. 8. Residual Histogram

Histogram shown in Figure 8 illustrates normal distribution and the model is best suited for accurate predictions. Evaluating the performance of the trained model is rendered by calculating the regression evaluation metric. The evaluation metrics extensively used in linear regression to validate the performances are :

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

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

Mean Squared Error (MSE) is the mean of the squared errors:

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

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \qquad (3)$$

The following evaluation metrics are calculated, which yielded minsized error values for each metric as shown in table 4.

TABLE IV. MINIMIZED ERROR METRICS

```
from sklearn import metrics
print ('MAE:  ' , metrics.mean absolute error ( y test, predictions) )
print ('MSE:  ' , metrics.mean squared error ( y test, predictions) )
print ('RMSE:  ' , np.sqrt ( metrics.mean absolute error
( y test, predictions) ) )
```

```
MAE: 7.22814865343

MSE: 79.813051651

RMSE: 8.93381506698
```

IV. CONCLUSION

The study uses sophisticated machine learning libraries to implement linear regression for forecasting the best profit making platform among product application and website. The following conclusions are drawn from the study:

1) Predicted values are more precise while the error is minimized. The validation is done with other set of accomplished results obtained by using another conventional technique for the same set of input values as shown in table 5. The procedures of implementation are well documented in several literatures and hence this paper evades the discussion on the same.

2) Use of advanced libraries present better features to arrive at precise values within a short interval of time.

3) Product application seems to yield better visibility and an increase in profit to about 25 % as compared to websites.

4) This model may be extrapolated for any e-commerce industry with local customization.

TABLE V. SVR VERSUS LINEAR REGRESSION PREDICTION

Errors	***Support Vector Regression(SVR)***	***Linear Regression (LR)***
MAE	7.22814865343	50.82389818465409
MSE	79.813051651	4895.242544790398
RMSE	8.93381506698	69.9660099247513

REFERENCES

[1] Hirose, Hideo, Yusuke Soejima, and Kei Hirose, “NNRMLR: A combined method of nearest neighbor regression and multiple linear regression,” In 2012 IIAI International Conference on Advanced Applied Informatics, pp. 351-356, IEEE, 2012.

[2] Wang, Dehua, Yujing Gao, and Zhiping Tian,“One-Variable Linear Regression Mathematical Model of Color Reading and Material Concentration Identification,” In 2017 International Conference on Smart City and Systems Engineering (ICSCSE), pp. 119-122, IEEE, 2017.

[3] Kavitha, S., S. Varuna, and R. Ramya, “A comparative analysis on linear regression and support vector regression,” In 2016 Online International Conference on Green Engineering and Technologies (IC-GET), pp. 1-5, IEEE, 2016.

[4] Bayindir, Ramazan, Murat Gok, Ersan Kabalci, and Orhan Kaplan, “An intelligent power factor correction approach based on linear regression and ridge regression methods,” In 2011 10th International Conference on Machine Learning and Applications and Workshops, vol. 2, pp. 313-315, IEEE, 2011.

[5] Men, Hong, Shanshan Zhang, Jiyong Jin, and Zhiming Xu, “Simultaneous determination of Pb and Cd ions with ion selective electrodes based on multiple linear regression,” In 2009 Third International Symposium on Intelligent Information Technology Application, vol. 1, pp. 415-418, IEEE, 2009.

[6] Guo, Liang, and Xianghui Deng, “Application of improved multiple linear regression method in oilfield output forecasting,” In 2009 International Conference on Information Management, Innovation Management and Industrial Engineering, vol. 1, pp. 133-136, IEEE, 2009.

[7] Harimurti, R., Y. Yamasari, and B. I. G. P. Asto, “Predicting student’s psychomotor domain on the vocational senior high school using linear regression,”In 2018 International Conference on Information and Communications Technology (ICOIACT), pp. 448-453. IEEE, 2018.

[8] Peng, Zhihao, and Xucheng Li, “Application of a Multi-factor Linear Regression Model for Stock Portfolio Optimization,” In 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), pp. 367-370, IEEE, 2018.

[9] Feng, YouLi, and ShanShan Wang, “A forecast for bicycle rental demand based on random forests and multiple linear regression.” In 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), pp. 101-105, IEEE, 2017.

[10] Naseem, A. Imran, B. Roberto Togneri, and C. Mohammed Bennamoun, “Face identification using linear regression,” In 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 4161-4164, IEEE, 2009.