ISSN No:-2456-2165

Predictive Analysis of Business Outsource of a Distribution Company using Machine Learning Techniques

Mandepudi Nobel Chowdary
Electronics and Computer Engineering
Sreenidhi Institute of Science and Technology
Hyderabad, India

Choudhary Vineet
Electronics and Communication Engineering
Sreenidhi Institute of Science and Technology
Hyderabad, India

Katta Rakesh
Electronics and Computer Engineering
Sreenidhi Institute of Science and Technology
Hyderabad, India

Chennupati Kumar Chowdary
Electronics and Communication Engineering
Sreenidhi Institute of Science and Technology
Hyderabad, India

Kontham Akhilesh
Electronics and Computer Engineering
Sreenidhi Institute of Science and Technology
Hyderabad, India

Abstract:- In the present scenario, every product based company need an analysis of the sales of the products sold across the various outlets located around the country or world. The business growth depends on the sales of the product to the customers with their satisfaction. In this way, every business distribution company needs the prediction to analyze the sales of the product across their showrooms. In this paper, the research study towards the analysis of the various data elements to predict the business outcome of a company is discussed with different machine learning models. Different attributes will play a key role to define the sales of the product that includes the factors of the customer or applicant, features of the product and the qualities of the managers in the store, who promote the product sales. The research to consider the important attributes, and the analysis of the data by exploring the data elements, the importance to impute the missing values plays a key role to increase the rate of perfect predictions and the model building using a different machine learning techniques for better accurate predictions are completely discussed in this paper that enhance to provide the better business sales predictions. This research study might be very much useful to the distribution companies, to consider the important key attributes that play a major role in the sales of their products.

Keywords:- *Prediction, Attributes and Business Outcome.*

I. LITERATURE SURVEY

Many companies mainly concentrate on the product and services and have to put growth on sales to promote the business. Every industry needs analysis of product sales. The rate of the sales of the product depends on the promotions made by the company and the rate of advertisements done by the company [1]. Out of the work from their side, the involvement of the sales managers to impress the customers and encourage them to buy the product. There will be different aspects like the previous sales done by the manager, his qualification, communication skills and many more to evaluate their performance [2]. The future analysis of the sales can be predicted based on the multiple varying factors related to the manager and the customer. The predictive modelling using a variety of machine learning algorithms to predict the future sale analysis of the product for a company will show better opportunities for further improvements [3]. The model can be designed with the optimum output by neglecting the error rate by considering the most important factors. There have been many companies, working at a better level with the future prediction of the sales. There will be about 10%-20% sales growth if future sales can be predicted [4]. The sales growth can be controlled based on the factors that are conflicting the decrement in the sales, which are identified with the model [5]. The predictive modelling will play a key role in the identification of the better modelling of the existing data to predict the future and reduce the errors. This research study will also be helpful to promote the sales of the distribution company.

ISSN No:-2456-2165

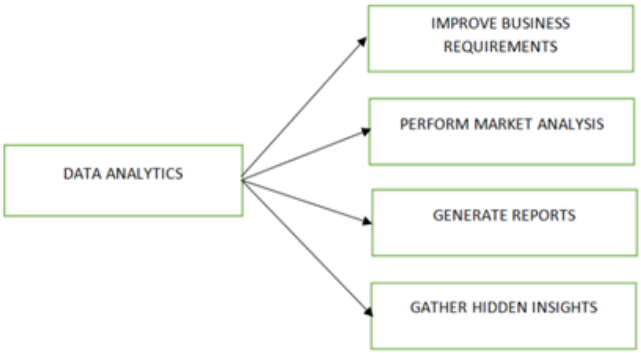
II. MODEL IMPLEMENTATION

In this research study, a dataset is taken into the consideration, showing certain details along with business outsourcing. The data elements are useful to work further in the model implementation.

A. Data Analysis

1) Data Analytics:

Data analytics refers to analyzing the data and improving their productivity and business gain. The techniques used for the analysis of the data varies from one organization to another organization and from one individual to another individual, it is also a process of inspecting, cleaning, transforming and modelling the data for the sake of understanding and achieving the required goals and concluding according to the requirements [6]. In this research study, the data of the product sales business outsource has to be analyzed adequately to ensure the better performance of the model.



2) Statistical Analysis:

It deals with the mathematical operations and collecting data and interpretation of data and presentation of data in different formats according to the requirements of the user or the organization to solve the complex problems. The data is properly collected with various data elements that play a role in business outsourcing. The hypothesis can also be generated based on the statistics, that the model will show better results or not with the existing data based on the statistics [7]. With the analysis of dataset different important data elements are;

1. Applicant _City_PIN
2. Applicant_Gender
3. Applicant_Martial_Status
4. Applicant_Occupation
5. Applicant_Qualification
6. Manager_Joining_Designation
7. Manager_Current_Designation
8. Manager_Grade
9. Manager_Status
10. Manager_Gender
11. Manager_Num_Application
12. Manager_Num_Coded
13. Manager_Business
14. Manager_Num_Products
15. Business_Sourced

3) Data cleaning:

The process of removing unwanted data and the process of removing inaccurate data from the available data set is known as the data cleaning which helps the user to stick to the required data only and can be reached to the conclusions. In this dataset to identify the better model results, the data cleaning has been done by removing the data elements Applicant_ID and Office_PIN.

B. Data exploration

1) Imputing missing values

The raw data which is not well organized and which may consist of the null values may mislead in the predictions of the results and also there will be the impact on the accuracy of the trained model with the missing values to overcome the consequences we will be inheriting some methods which help in increasing the accuracy and a well-predicting model.

Generally, we follow this mechanism when we have more than one data set and we least bothered about one sample data so we use deleting records of missing value technique. Other than that we can create a separate model without any missing values, but it takes good enough time to remove the records with the missing values [8]. This method of the new dataset would be helpful to build the model with better performance. Although these techniques will not work with the dataset we are dealing with. Hence, the statistical methods are useful to calculate mean, mode, median, variance and standard deviation. The missing values in the dataset can be replaced with the most suitable statistical calculation [9].

From our dataset, it can be observed that there are many missing elements, especially in some particular datasets. To ensure the better performance of the model, we must ensure to impute those with a suitable statistical method.

```
ID                                0
Office_PIN                        0
Applicant_City_PIN                0
Applicant_Gender                 53
Applicant_Marital_Status         59
Applicant_Occupation           1090
Applicant_Qualification          71
Manager_Joining_Designation       0
Manager_Current_Designation       0
Manager_Grade                     0
Manager_Status                    0
Manager_Gender                    0
Manager_Num_Application           0
Manager_Num_Coded                 0
Manager_Business                  0
Manager_Num_Products              0
Manager_Business2                 0
Manager_Num_Products2             0
Business_Sourced                  0
dtype: int64
```

ISSN No:-2456-2165

In the dataset, we can find the fields of Applicant_ City_ PIN, Applicant_ Gender, Applicant_ Martial_ Status and Applicant_ Occupation with major missing values. We can impute these fields with the mode of the specific field, thus the missing values are replaced with the most frequently occurring value. The missing values in the Applicant_Gender are filled with the 'M', which is the most repeated value in this particular field.

```
df['Applicant_Gender'].fillna('M', inplace=True)
```

The missing values in the Martial_Status can also be imputed with the 'M', as most of the fields are with a similar type.

```
df['Applicant_Marital_Status'].fillna('M', inplace=True)
```

The Applicant_Occupation can also be imputed with the 'Salaried' in the missing tuples, as most of the applicants have a similar one.

```
df['Applicant_Occupation'].fillna('Salaried', inplace=True)
```

Finally, the Applicants_Qualification can be imputed with the 'CLASS XII', in the missing areas as it is the statistical mode value.

```
df['Applicant_Qualification'].fillna('Class XII', inplace=True)
```

Using this method we can get more accurate results and the time factor is a benefit in this technique. We can conclude with all the results rather than neglecting the missing values from the sample data and thus we can obtain the dataset with imputing the missing values, suitable to design the model.

III. MODEL BUILDING USING LOGISTIC REGRESSION

Regression models are one of the most popular statistical models (or techniques) used for predictive modelling and data mining. Through regression models, we can be able to predict the target variables (either continuous or categorical). Many of the data scientists use regression models for solving their problem statements. The two types of regression are:

Linear regression is mainly used when there is a relationship between dependent and independent variables. Here, the increase/decrease in one variable leads to an increase/decrease of the other.

Logistic regression is a supervised machine learning algorithm (a machine learning algorithm where the target variable is present) mainly used to predict the probability of a target variable. Here the outcome is categorical (i.e. the target is yes/no) [10].

For our dataset logistic regression is very much suitable, where the target variable to be predicted is clearly defined with the required data elements that are to be trained [11]. Initially, the list of the categorical variables in the dataset has to be identified before training the model. The categorical variables are 'Applicant_Gender', 'Applicant_Martial_status', 'Applicant_Occupation', 'Applicant Qualification', 'Manager_Joining_Designation', 'Manager_Current_Designation', 'Manager_Status', 'Manager_Gender'.

Now, the data has to be split into the two parts, one part is to be trained to the model and the other to be tested which is useful to judge the prediction rate of the trained model. After the successful training of the data to the regression model, the roc score factor can be obtained to judge the prediction rate. For our data model, with this logistic regression, the prediction rate is about 46.9%, which is quite low. Thus this model's predictions cannot be considered to work further with such a low prediction rate.

```
roc_auc_score(valid_y, pred_valid[:,1])
```

```
0.4697613206972208
```

Hence the data fit the logistic regression with a wide range of data elements is not properly done. Another machine learning model of the random forest will be better to predict with better performance rate with larger data elements.

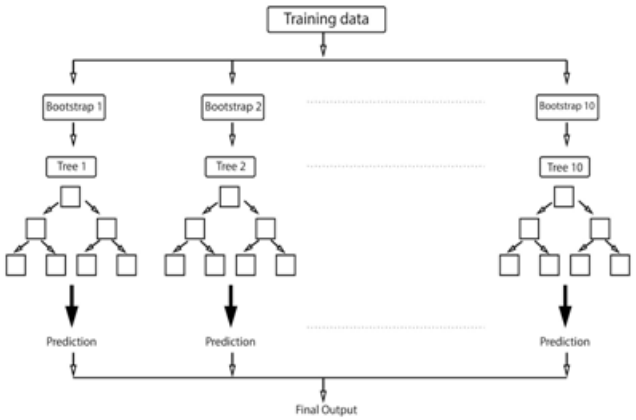
IV. MODEL BUILDING USING RANDOM FOREST CLASSIFIER

Random forest is an ensemble learning algorithm used for both classification and regression problems. It is mainly constructed through decision trees. First, we should create a bootstrap sampling for the construction of a Random Forest. Bootstrap sampling is a resampling technique used to estimate statistics (like population) by sampling a dataset. It is mainly used in machine learning algorithms to predict the performance of the data model which is in the test data set Simply we can say that bootstrap sampling means creating multiple samples of the same size from the original dataset. Now, we should create our models for those multiple samples and ensemble them to get the right prediction [12]. Through these bootstrap samplings, we should create a random forest for getting a better prediction for our test dataset.

➢ *Steps for implementing a Random Forest*

- Create multiple bootstrap samples
- Build a decision tree on every sample
- Use feature sampling for each split in the decision tree
- Aggregate all decision trees
- Combine all bootstrap samples to make a random forest

The train data can be created into multiple samples by ensembling them into different bootstrap samples using the random forest classifier. This way of creating separate samples into the multiple sets and training the different models will reduce the chances of the error and ensure better performance. This process of separate samples will able to train the vast data elements properly.



➢ *Hyper Parameters of Random Forest*

1. max_features
2. max_depth
3. min_samples_split
4. min_samples_leaf
5. Criterion
6. n_estimators (no of trees)

For better model performance the number of trees should be less as possible. We can select the best hyperparameters for measuring the performance of the model Repetition of models is allowed in bootstrap sampling [13]. Max_depth is a stopping criterion for random forest models. With all these factors, the training data is fit to the random forest classifier accordingly with the suitable values.

```
rf.fit(train_x, train_y)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=4, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=2, verbose=0,
                       warm_start=False)
```

With this trained model using the random forest classifier, the roc score factor is about .57, which shows that the prediction rate has been increased to the value of 57%, which is very much better compared to the above logistic regression model outcome.

```
roc_auc_score(valid_y, pred_valid_rf[:,1])
```

```
0.5707349946164473
```

V. RESULTS

The above predictive data modelling using different machine learning techniques show a wide variety in the prediction rate. The first model with a large number of data elements shows the less efficient output prediction, whereas the second model with samples trained as the models using the random forest classification shows the better output prediction rate, which is quite good. The prediction rate of the same data with the first model is 46.9%, which is drastically increased to 57% with the second model using different machine learning models. Even the prediction factor in the second model can be increased to some good level with the better imputing of the missing values and consideration of large data [14]. Thus the results of the research study show that the prediction of the business outsource of the distribution company can be predicted with 57% confidence. Thus the result ensures that this study might be useful in the further improvements of the prediction rates and types [15].

VI. CONCLUSION

This paper shows the research study on the sales of the product can be very much useful to analyze the future and make wise decisions to drive the better way in the business [16]. This paper will be useful for further predictions on the various business promotions to enhance the better goals in the future. This sort of prediction on every factor responsible for the growth of the business will make the company get to know about the needs and quality that is being expected from the customers [17] [18]. Therefore on the commercial aspect, the entire research study with the optimum outputs obtained will be useful to the different companies to increase their sales and promote their business.

REFERENCES

[1]. Goedhuys, M. and Veugelers, R., 2012. Innovation strategies, process and product innovations and growth: Firm-level evidence from Brazil. *Structural change and economic dynamics*, *23*(4), pp.516-529.

[2]. Giering, M., 2008. Retail sales prediction and item recommendations using customer demographics at store level. *ACM SIGKDD Explorations Newsletter*, *10*(2), pp.84-89.

[3]. Yaseen, R.M., Shah, H.J., Pavlov, A.L., Manjunath, J. and Bhatia, M., Oracle International Corp, 2012. *Sales prediction and recommendation system*. U.S. Patent Application 13/236,629.

[4]. Ribeiro, A., Seruca, I. and Durão, N., 2016, June. Sales prediction for a pharmaceutical distribution company: A data mining based approach. In 2016 11th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-7). IEEE.

[5]. Mann, L., Samson, D. and Dow, D., 1998. A field experiment on the effects of benchmarking and goal setting on company sales performance. Journal of Management, 24(1), pp.73-96.
[6]. Weir, B.S., 1990. Genetic data analysis. Methods for discrete population genetic data. Sinauer Associates, Inc. Publishers.
[7]. Brandt, S., 1976. Statistical and computational methods in data analysis (No. 04). Amsterdam, The Netherlands:: North-Holland Publishing Company.
[8]. Royston, P., 2004. Multiple imputation of missing values. The Stata Journal, 4(3), pp.227-241.
[9]. Acock, A.C., 2005. Working with missing values. Journal of Marriage and family, 67(4), pp.1012-1028.
[10]. Peng, C.Y.J., Lee, K.L. and Ingersoll, G.M., 2002. An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), pp.3-14.
[11]. Wong, G.Y. and Mason, W.M., 1985. The hierarchical logistic regression model for multilevel analysis. Journal of the American Statistical Association, 80(391), pp.513-524.
[12]. Khaidem, L., Saha, S. and Dey, S.R., 2016. Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.
[13]. Cootes, T.F., Ionita, M.C., Lindner, C. and Sauer, P., 2012, October. Robust and accurate shape model fitting using random forest regression voting. In European Conference on Computer Vision (pp. 278-291). Springer, Berlin, Heidelberg.
[14]. Grabowski, H.G. and Vernon, J., 2000. The distribution of sales revenues from pharmaceutical innovation. Pharmacoeconomics, 18(1), pp.21-32.
[15]. Aghdaie, M.H., Zolfani, S.H. and Zavadskas, E.K., 2014. Sales branches performance evaluation: a multiple attribute decision making approach. In 8th International Scientific Conference "Business and Management 2014 (pp. 1-7).
[16]. Bohanec, M., Robnik-Šikonja, M. and Borštnar, M.K., 2017. Organizational learning supported by machine learning models coupled with general explanation methods: A Case of B2B sales forecasting. Organizacija, 50(3), pp.217-233.
[17]. Khalil Zadeh, N., Sepehri, M.M. and Farvaresh, H., 2014. Intelligent sales prediction for pharmaceutical distribution companies: A data mining based approach. Mathematical Problems in Engineering, 2014.
[18]. Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., Chu, S. and Yang, X., 2015, January. On machine learning towards predictive sales pipeline analytics. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (pp. 1945-1951).