Customer Centric Sales Analysis and Prediction

Shiwani Joshi, Lavi Samuel Rao, B. Ida Seraphim

Abstract*—For successful business management, sales prediction plays an inevitable part. Data mining techniques have been employed since a long time for sales analysis. In the past, the prediction has been done from various point of views always keeping mind the needs of the customer and the profitability of the business in consideration. Initially, sales prediction has been done using Market Basket analysis, wherein using the previous data the next item , which is most likely to be purchased is predicted. At later stages, the products on the shelf were only considered for predicting sales in a supermarket. Thereafter , for a range of supermarkets the location and time was considered to predict the sales of items. For predicting the sales a number of algorithms such as AIS, Apriori, FP Growth, FP Bonsai have been employed. In this paper, the price or the amount to be likely spent by the customer will be predicted using various other algorithms and a comparison between the different algorithms is outlined.

I. INTRODUCTION

Data mining is the field for examining pre-existing database for excavating new information. This field forms the basis of Analytics and is used to make predictions for various field such as marketing, finance, medicare, weather, etc. Many times the term data mining is construed to finding relevant data from the pre-existing data set instead, it is actually the extraction of relevant information or pattern for making predictions. Data mining is the field which is inevitably used in the field of retail and marketing, and when retail comes into question sales is the term that is coined next. The business of Sales can work more efficiently when data mining techniques are inculcated in it. In the process of prediction , the initial requirement is to have a clear data set. The data which we get needs to be cleaned up first by using techniques like machine learning, statistics, etc. The field of data mining is not only restricted to retail and sales but has numerous other applications too which are listed as follows:

- Analysis in Finance sector
- Intrusion Detection
- Health sector
- Retail and marketing
- Scientific Approaches for research purpose

Manuscript published on 30 April 2019.
***** Correspondence Author (s)

Shiwani Joshi, Computer Science and Engineering , SRM Institute of Science and Technology, Chennai, India.
Lavi Samuel Rao, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
B. Ida Seraphim, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

We have implemented the techniques of data mining basically in the field of retail and marketing. The application of data mining for Black Friday Sales will not only benefit the customers but also the providers. By getting a clear insight as to what a particular section of consumer is likely to spend on what category of products can act as a game changer for the business. This process of analysis and prediction of sales can be done using various data mining techniques. Similar work, has also been done on Chinese tobacco sales prediction[1] wherein a data mining technique was used to predict the sales along with neural net model and a comparison between the two was also done. For prediction of amount that would be spent by any customer, data mining is particularly called into play since it involves mining of certain patterns which can be done by following the steps of KDD(Knowledge Discovery in Databases) process. The steps are listed as follows:

1) Selection of Data set
2) Pre-processing
3) Transformation
4) Data Mining
5) Interpretation

The most important and significant step of data mining is to select the best available data set. Once the data set is selected, it is suppose to be made free of outliers and noise by pre-processing and transformed into a form in which it is needed for getting mined. A better insight of the data set can be obtained by visualization of the various attributes.

Data mining techniques have been employed on Women Ap- parel dataset, tobacco dataset, non-alcoholic beverages dataset, Supermarket dataset are worked upon to predict the sales.

The simplified steps for prediction of sales is as follows:

- Pre-processing
- Data Mining
- Validation of results

Pre-processing:

After obtaining a good data set, pre-processing of the data set needs to be done so that outliers and noisy data could be gotten rid of. For pre-processing , a number of techniques are used such as one hot encoding, glove vectorisation, up scaling, etc. Visualization of the data set could help a great deal in understanding the data set and the relationship of the various attributes with one another. The pre processing stage involves steps like data augmentation, cleaning of the data, etc. Visualization also plays an important role in understanding and creating a relationship between various attributes.

In other words, visualization is to help understand by efforts of visual effects using charts, graphs, histograms, etc.

Data Mining:

Data mining is the process of finding the suitable pattern from the data set available. These patterns are excavated by means of intersection of machine learning, statistical and database system techniques. This step deals mainly with the following classes of tasks:

1) Association rule mining-It involves the mining of association rules using dependency among variables. For example Market Basket Analysis wherein on the basis of the transactions the probability of buying an item "X" on account of buying item "Y" is predicted.
2) Regression-In statistical modeling , regression analysis is used to find relations between dependent and inde- pendent variable.
3) Clustering-It involves the grouping of "similar" items together in form of cluster using algorithms like K- means clustering.
4) Classification-It is used to predict categorical class labels by means of various models like naive Bayes, decision tree.
5) Anomaly Detection-Data mining techniques are also used to detect unusual patterns, which do not conform to the expected results called outliers. Support Vector Based Anomaly detection technique, Clustering Based Anomaly Detection Technique, etc.
6) Summarization-This task helps to give the user a better and compact information about the data and the most common tool used for the same is Excel.

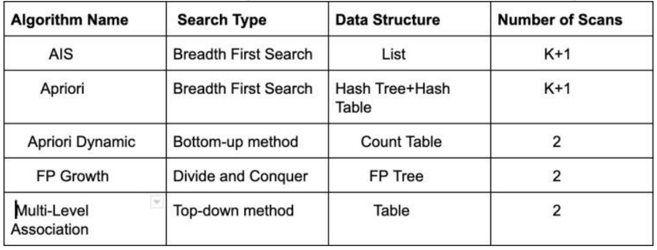
Validation of Result:

Generally , the results are either in the form of class labels in case of classification or predict the dependency of one variable on another. The results that are mined are not necessarily according to what we need. In order to check for the accuracy of the results or patterns mined, we must apply certain tests which could predict the accuracy of the models or methods applied. Therefore, the final step is to validate the results by using methods like the ROC curve, accuracy , precision of the resultant. There might be certain errors in the used model like over fitting or under fitting. Certain measures would be needed to be taken to overcome these errors before arriving on the final conclusion so that accurate results could be obtained.

II. LITERATURE SURVEY

Until now, work has been done in this field, beginning with the market basket analysis [1] using Association rule mining. A number of other algorithms like Apriori [2] have been called into play for Market Basket Analysis. A lot of work has been done for finding association rules in a large group of database using more advanced frameworks [3] [4] [5] [6] [7] . In all these algorithms A number of new algorithms like CLASD [8] also came into light for getting better results using small local- ized segments. Another algorithm, ROCK [9] , also came into existence for market basket analysis. A lot of work has been done in the multi-store environment as well.In comparison to the traditional Apriori algorithm, Apriori-like algorithm [10] was used for a multi-store environment. This algorithm also considered the time and location of the store for each item's PT table. This algorithm proved to be computationally better as compared to the traditional Apriori algorithm. Although the Apriori algorithm was among the pioneer algorithms of this field but it has had several drawbacks which don't make it the best choice [11] . The Apriori algorithm does several scans of the database and a lot work increases.In addition to the Apriori algorithm, other algorithms like AIS, FP Growth, Apriori Dynamic Programming and Multi-level Association have also been worked with. A comparative study of the same has been made [12] and shown in figure 1.

In [12] multi-level association, different levels of abstractions are used for mining frequent patterns. Another algorithm brought into light was the FP-Growth algorithm [13] [14] wherein a tree structure is used for two-step approach towards mining useful patterns. This approach also has certain disadvantages of not being of any use in the incremental mining process. Until now, only historical database has been considered but there has to be a way where even the current situations are also considered as in case of [15] .In [15] , Recommender System is used wherein four step process is used:



Algorithm Name	Search Type	Data Structure	Number of Scans
AIS	Breadth First Search	List	K+1
Apriori	Breadth First Search	Hash Tree+Hash Table	K+1
Apriori Dynamic	Bottom-up method	Count Table	2
FP Growth	Divide and Conquer	FP Tree	2
Multi-Level Association	Top-down method	Table	2

Fig. 1. Comparative study

- Retrieve the most relevant case
- Reuse the knowledge provided in the next case
- Revise the solution obtained
- Retain the parts of the new solution that are likely to be used for future purpose.

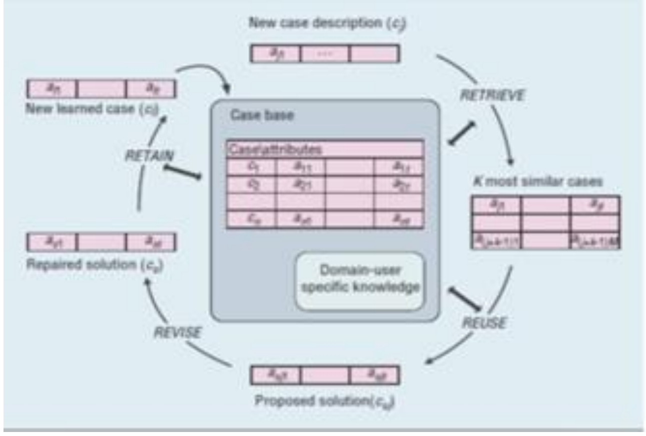


Fig. 2. The CBR cycle consisting of four steps

The implementation has been done using the findings given in [15] .

III. METHODOLOGY

A. Dataset

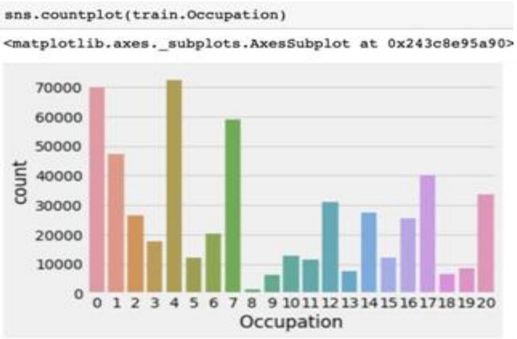
The data set has been taken from Analyticsvidya.com and it comprises twelve attributes and a total of 550068 entries. The attributes are the various characteristics of the customers like the age, city of residence, occupation, marital status, number of years of residence and there are also attributes about the product like the product ID and the product category.The data set is huge and had a missing value as well. Dealing withsuch a huge data set is itself a very tedious job and hence a number of techniques were needed to be applied during the pre-processing step.

B. Data Visualization

The very initial step in examining the data set is data visualization. Data visualization is done using the graphical representation of the relationship of various attributes with each other. A lot of inferences have been made at the data visualization stage like the number of null entries and the attributes having null entries. There have been graphs made to establish relationship between the target variable(Purchase) and the other attributes.Broadly, there have been two types of analysis done namely, univariate and bivariate analysis. The univariate analysis is the graphical representation of single attributes which gives us the total count of the attribute like number of people performing a certain type of job or number of female, male customers.

1. Exploratory Data Analysis (EDA)

Fig. 3. Univariate Analysis

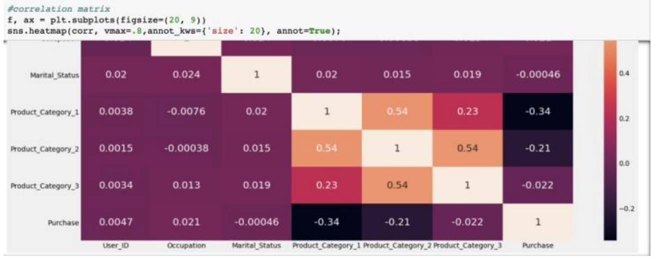


```
sns.countplot(train.Occupation)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x243c8e95a90>
```

Fig. 4. Occupation count

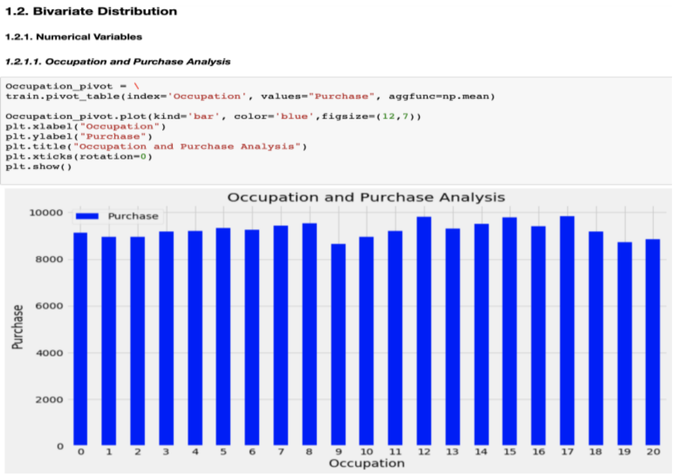
Correlation analysis was also done to know the degree of dependency of one attribute on the other attributes. The correlation analysis shows that the product category 1, product category 2 and product category 3 are somewhat dependent on each other unlike the other attributes. This might be so because certain products maybe categorized as product category 1 as well as product category 2.



```
#correlation matrix
f, ax = plt.subplots(figsize=(20, 9))
sns.heatmap(corr, vmax=.8,annot_kws={'size': 20}, annot=True);
```

	User_ID	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
Marital_Status	0.02	0.024	1	0.02	0.015	0.019	-0.00046
Product_Category_1	0.0038	-0.0076	0.02	1	0.54	0.23	-0.34
Product_Category_2	0.0015	-0.00038	0.015	0.54	1	0.54	-0.21
Product_Category_3	0.0034	0.013	0.019	0.23	0.54	1	-0.022
Purchase	0.0047	0.021	-0.00046	-0.34	-0.21	-0.022	1

Fig. 5. Correlation Analysis



1.2. Bivariate Distribution

1.2.1. Numerical Variables

1.2.1.1. Occupation and Purchase Analysis

```
Occupation_pivot = \
train.pivot_table(index='Occupation', values="Purchase", aggfunc=np.mean)

Occupation_pivot.plot(kind='bar', color='blue',figsize=(12,7))
plt.xlabel("Occupation")
plt.ylabel("Purchase")
plt.title("Occupation and Purchase Analysis")
plt.xticks(rotation=0)
plt.show()
```

Fig. 6. Bivariate Analysis

C.Pre-Processing

In pre-processing, the data set is checked for null values and those values are replaced with random values.In this dataset null values were found in the attributes Product_Category 2 and Product_Category 3, which were replaced with the random digit -2.0. Thereafter, group 19 and 20 were removed from Product_Category 1 so that equal number of groups are there in all the product categories.

It would further help in better analysis and model building. The stage of pre-processing also involved the conversion of categorical data to numerical data so that the modeling is easier. The categorical data for attribute Gender, Occupation, City_Category was changed to numerical values.

D. Model Building

This is the final stage of the project wherein different models were tried for getting accurate prediction of the target variable "Purchase". The testing data comprised of 233599 entries and the training data had 550068 entries.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

A number of machine learning algorithms were tried to get an accurate prediction. The performance measure used is the Root Mean Squared Error(RMSE) wherein the lesser this value the better

is the model under consideration. The RMSE value shows that how different is the predicted outcome from the expected outcome. Five models were used to get the best result possible namely, Linear Regression, Ridge Regression, Random Forest, Decision Tree and XG Boost. The RMSE value of the models are calculated so that the best model is considered.

1) *Linear Regression Model*

The linear regression model is one of the basic machine learning algorithms considered for regression analysis. The linear regression model and data is taken as the input for the generic function that is used to build the model. The model coefficients are plotted to analyze the effect of attributes on the prediction. For linear regression the attribute that has an effect on the prediction is Stay_In_Current_City_Years. The RMSE value obtained for Linear regression is 0.4339 which is very high.

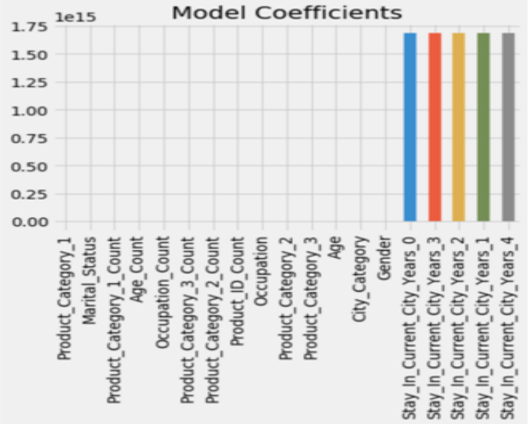


Fig. 7. Model coefficients for Linear Regression

2. *Ridge Regression Model*

The ridge regression model used for prediction is different from linear regression and performs better because in ridge regression the value of alpha can be set manually and the best value is chosen. In this case the value of alpha chosen is 0.05 which gives the best results and accurate predictions. The RMSE value which is 4346 for this model that has not improved.It might be because of the tediousness of the improved model.

3. *Decision Tree Model*

The decision tree model is considered for prediction of the target variable. But using this model also only a few attributes have an impact in predicting the target variable. Although the RMSE value for this model has improved which shows

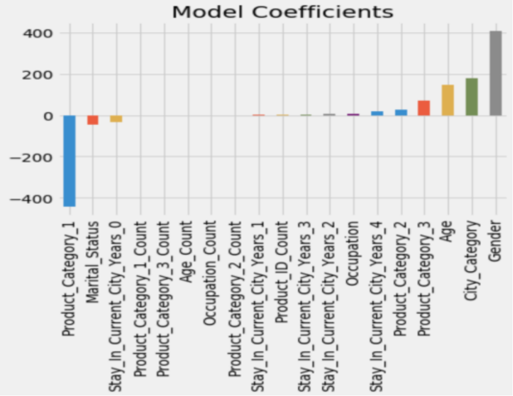


Fig. 8. Model coefficients for Ridge Regression

that the predicted outcome is somewhat same as the expected outcome. The RMSE value for the decision tree model is 2680.

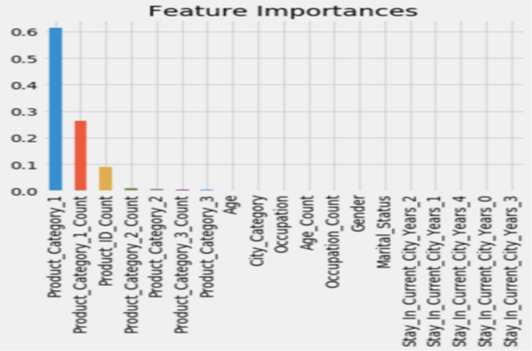


Fig. 9. Feature importance for Decision Tree

4. *Random Forest Model*

The random forest model is a collection of decision trees. It is used for preventing over fitting of the model since a single deep decision tree could result in over fitting. It is done by forming random subsets of the features and shallower trees. The RMSE value using this model has also improved and has come out to be 2803.

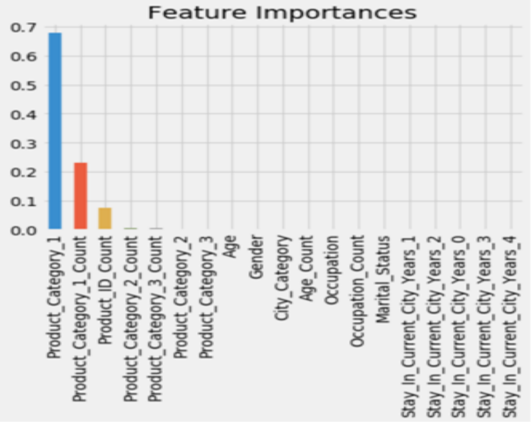


Fig. 10. Feature importance for Random Forest

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

5. XGBoost Model

This model is considered the most advanced one as of now and is used to get an accurate result. The RMSE value for the same came to be 2815.

IV. CONCLUSION

The field of sales and marketing plays vital role in enhancing one's business. The work that has been done until now only considered the past records and predict the next item that is most likely to be purchased by the customer. In the multi store environment, the time and store location have also been considered for improving the sales. But if the purchase amount could be predicted as done here then the store could target a particular section of the customers and improve the profit margin for the store. After prediction of the purchase amount, certain marketing strategies could be applied for certain section of the customers so that the profit of the store could be enhanced.

REFERENCES

1. R. Srikant and R. Agarwal, "Fast Algorithms for mining Association rules," 1994, pp. 478–499.
2. R.Agrawal, T.Imielinski, and A.Swami, "Mining association rules between sets of items in large databases," 1993, pp. 207–216.
3. M. Klementtinen, P. Ronkainen, H. Toivonen, and H. Mannila, "Finding Interesting Rules from Large Sets of Discovered," 1994, pp. 401–407.
4. G. Piatetsky-Shapiro, P. Smyth, and U. M. Fayyad, "From Data Mining to Knowledge Discovery: An Overview of Advances in Knowledge Discovery and Data Mining," 1996, pp. 1–34.
5. R.M.C.Silverstein and S.Brin, "Beyond Market Baskets: Generalizing Association Rules," in *ACM SIGMOD Conf*, 1997.
6. C. C. Aggarwal and P. S. Yu, "A New Framework for Itemset Generation," in *ACM PODS Conf*, 1998.
7. S. S. S. Chakrabarti and B. Dom, "Mining Surprising Patterns Using Temporal Description," in *Length of Proc. Int'l Conf. Very Large Databases*, 1999.
8. C. C. Aggarwal, P. S. Yu, and C. Procopiuc, "Finding Localised Associations in Market Basket Data," *IEEE Transaction on Knowledge and Data Engineering*, vol. 14, 2001.
9. R. R. S. Guha and K. Shim, " ROCK: A Robust Clustering Algorithm for Categorical Attributes," 1999.
10. Y. L. Chen, K. Tang, R. J. Shen, and Y.-H. Hu, "Market Basket Analysis in a multiple store environment," 2004, pp. 339–354.
11. S. Gupta and R. Mamtora, "A survey on Association Rule Mining in Market Basket Analysis," *International Journal of Information and Computational Technology*, pp. 409–414, 2014.
12. D. Fol and P. C. Chaudhary, "Finding an Efficient Approach for Generating Frequent Patterns in Large Database," pp. 20–27, 2015.
13. J. Guo, P. Zhang, J. Tan, and L. Guo, "Mining Hot Topics from Twitter Streams," in *International Conference on Computational Science*. El- sevier, 2012.
14. M. S. B. Phridviraja and C. V. Gururao, "Data mining–past, present and future–a typical survey on data streams." Elsevier, 2013.
15. A. Gatzioura and M. Sànchez-Marrè, "A Case - Based Recommender Approach for Market Basket Data," *IEEE Intelligent Systems*, 2015.