BUY AND SELL TREND ANALYSIS USING MACHINE LEARNING AND DEEP LEARNING

A major project report submitted in partial fulfillment of

the requirement for the award of degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Lakshmikanth Loya (221710309032)

Shivva Saiteja (221710309056)

Jampani Kranthi Kumar (221710309018)

Kandey Jeevan Kumar (221710309022)

Under the esteemed guidance of

Dr. G Himabindu

Asst. Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GITAM

(Deemed to be University)

HYDERABAD

MAY - 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GITAM SCHOOL OF TECHNOLOGY

GITAM

(Deemed to be University)

DECLARATION

We, hereby declare that the major project report entitled **"BUY AND SELL TREND ANALYSIS USING MACHINE LEARNING AND DEEP LEARNING"** is an original work done in the Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University) submitted in partial fulfillment of the requirements for the award of the degree of "Bachelor of Technology" in Computer Science and Engineering. The work has not been submitted to any other college or University for the award of any degree or diploma.

Date:

Lakshmikanth Loya (221710309032)

Shivva Saiteja (221710309056)

Jampani Kranthi Kumar (221710309018)

Kandey Jeevan Kumar (221710309022)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GITAM SCHOOL OF TECHNOLOGY

GITAM
(Deemed to be University)

CERTIFICATE

This is to certify that the project report entitled **“**BUY AND SELL TREND ANALYSIS USING MACHINE LEARNING AND DEEP LEARNING**”** is a bonafide record of work carried out by **Lakshmikanth Loya(221710309032), Shivva Saiteja(221710309056), Jampani Kranthi Kumar(221710309018), Kander Jeevan Kumar (2217103)** submitted in partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering.

Project Guide	**Head of the Department**
Dr. G Himabindu	**Dr. S. Phani Kumar**
Assistant Professor	**Professor**

ACKNOWLEDGMENT

Our project would not have been successful without the help of several people. We would like to thank the personalities who were part of our project in numerous ways, those who gave us outstanding support from the birth of the project.

We are extremely thankful to our honorable Pro-Vice Chancellor, **Prof. N. Siva Prasad** for providing necessary infrastructure and resources for the accomplishment of our project.

We are highly indebted to **Prof. N. Seetharamaiah**, Principal, School of Technology, for his support during the tenure of the project.

We are very much obliged to our beloved **Dr. S. Phani Kumar**, Head of the Department of Computer Science & Engineering for providing the opportunity to undertake this project and encouragement in completion of this project.

We hereby wish to express our deep sense of gratitude to **Dr. G Himabindu , Asst. Professor**, Department of Computer Science and Engineering, School of Technology for the esteemed guidance, moral support and invaluable advice provided by him for the success of the project.

We are also thankful to all the staff members of Computer Science and Engineering department who have co-operated in making our project a success. We would like to thank all our parents and friends who extended their help, encouragement and moral support either directly or indirectly in our project work.

Sincerely,

Lakshmikanth Loya (221710309032)

Shivva Saiteja (221710309056)

Jampani Kranthi Kumar (221710309018)

Kandey Jeevan Kumar (221710309022)

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

1.ABSTRACT

The uncertainty of the market conveys inherent risks for business. Some small-scale organisations with their limited resources require access to Artificial Intelligence for analysing their product trends, sales, customer behaviour etc. frequently. These steps are taken by the small-scale businesses in order to avoid being affected by the highly volatility and variance of the market. This current work-in-art paper was originally created to design a model that helps the small-scale businesses to analyse and predict the trendiness of a product at a reasonable cost using the concepts of machine learning known as decision trees. The limitations that we observed in this was - From time to time, the complexity of the tree was being reduced due to post-pruning and hence some cases with a greater number of features were being left out and could not be performed automatically by Python (Version-3.6) and Schitlearn. To overcome this limitation, we plan to compare the Decision Tree results to statistical machine learning algorithms like SVM and ANN (deep learning models included) and Gradient Boosting in terms of Accuracy and AUC curve.

Keywords: -Machine learning, Deep Learning, Market Trend Analysis,

2.INTRODUCTION

There are fluctuations in market on daily basis. Long term fluctuations and seasonal changes are estimated by business owner but can only do so if they have experience. The regular based fluctuations in market that pose a risk to business cannot be estimated. To minimize this risk or to handle such risky scenarios, businesses need to be prepared based on accounting and optimizing for possible outcomes. The field of Machine Learning and Data Science have introduced the concepts where the market features are taken into consideration and predictions are made to simplify the decisions to reduce risk.

Many machine learning techniques have been successfully implemented to predict the market trend. Different models give different performance, but the main goal is to predict the market trend accurately with less computational complexity. Although, Artificial neural networks have also been considered and their concept of trying to mimic brain is good for learning, there is a requirement of huge dataset for the model to be trained properly.

Decision trees have also been implemented on market trend and as decision tree are comparatively less computationally complex, the model suits the problem. However, as the market trend directly impacts the profit of business or organization the results being given by the model need to be very accurate.

We have considered boosting algorithms such as Extreme Gradient Boosting Classifier for our problem and it was also observed that the predictions made by Extreme Gradient Boosting Classifier are more accurate. These predictions are used by business vendors to take suitable steps and decrease the risk posed on the business. This work aims to provide business vendor a machine learning model that can predict the market trend more accurately and as the businesses are not noticeably big in scale, the model should be able to work with comparatively smaller dataset and is less computational complex.

3. LITERATURE SURVEY

As suggested by Carlos Vaca, et al, (2020) [2] we have used the three data sets demonstrating the market trends of different kinds. The author has used decision tree algorithms C 4.5 and CART variants to obtain the best results possible with less computational complexity. They have narrowed down the hyperparameters of the decision model to improve results. Using the work of the author we have improved the model through boosting.

Devpriya Soni, et al, (2018) [4] in their research have said that market trend analysis can be defined in a particular that is universally accepted. They say that different works have been done on analysis of market trend such as timeseries analysis, statistical analysis, etc. However, due to the influence of external factors playing a huge role in changes in trend no particular can be accepted universally. The author states that feature engineering and nature-based analysis can be used more an analysis momentarily which must later be updated.

Mojtaba Nabipour, et al, (2020) [7] has shown in his work of stock market prediction that by having huge datasets makes deep learning model a good choice. The author has compared over nine machine learning algorithms on market trend analysis and have considered Chinese stock market data of two years. In their work they have observed that deep learning model gave the highest results.

4. PROBLEM IDENTIFICATION AND OBJECTIVES

The main aim of the project is to analyze the trend of market and create a model that is less computational complex and gives good results at predicting the uncertain market trend. This will help the business vendors to take necessary steps to protect their business or reduce the risk from changing market trend.

For the purpose of making a model with less cost, the author here uses the concepts of "Decision Trees". A decision tree is a tool that supports decision by making tree like model of their possible sequences, including factors like chance event outcomes, resource costs, utility etc. In this approach the author uses two types of subsections that is C4.5 and CART. In C4.5, this tree improves its predecessor ID3 by normalizing the information gain measure and thus avoiding any biased to attributes with high values. As any DT, the construction of this tree depends on selecting the "best" attribute to split a dataset.

Whereas in CART which is the short form for Classification and Regression Trees, as the name suggests, CARTs are binary trees that use Gini Impurity as its attribute selection measure.

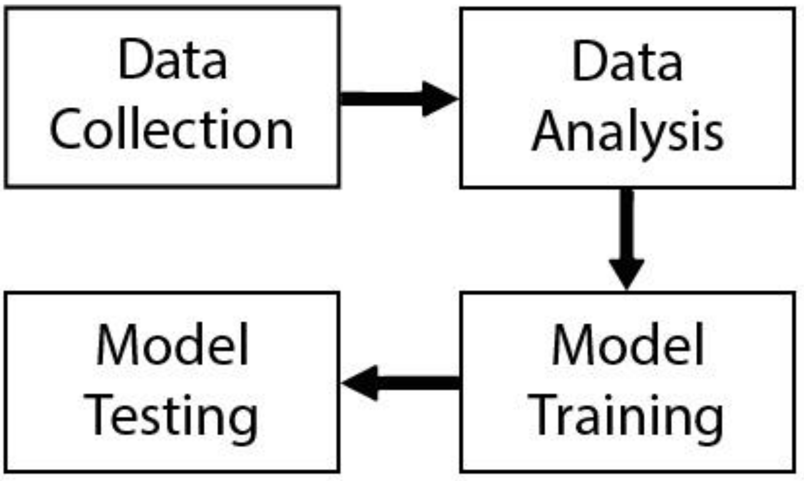
The idea behind this measure is that at any node of the tree a decision is made according to the least impurity.

Drawbacks

The limitations observed in this paper were that since the decision tree was simple due to post-pruning a lot of complex cases were being left out or ignored completely as the complexity of the tree was reduced. Post pruning in decision tree is a concept where we first build the decision tree and then remove the unwanted branches.

5. SYSTEM METHADOLOGY

Figure 5.1: Block diagram of System Architecture



We have four important phases in our project. The first phase is data collection. In data collection we gather the necessary and required datasets. For our work we have collect three different datasets. After data collection we go to data analysis phase. In this phase we analyze the data we have collected. We make visual representations of data for better and easy understanding. Also, through analysis we find the hidden patterns in data. After the analysis of data, we send the data for model training. In model training we first convert the categorical values into numeric values as machine learning models can only operate on numeric data. We train different models with all three different datasets. Finally, we go to the Model testing phase where we find the accuracy of predictions made by our models and compare the results.

6.OVERVIEW OF TECHNOLOGIES

Python

Python is an object-oriented and interpreted high level programming language which is used for general coding purposes. Its main goal is to spotlight code reusability with the use of significant indentations. Thanks to its object-oriented behavior, it helps us to write a clear and logical code for both small-scale and large-scale projects.

The main advantage of python is that it has many pre-defined libraries and functions where they can be imported directly which are used in various applications like databases, retrieving , processing and visualizing data, accessing web data etc.

Google Colab

Google Collaboratory is a one of the items from the catalogues of google research which allows everybody to design and execute a python code through a browser. It is helpful for us for implementing codes which involves concepts like Machine Learning, AI, Neural Networks with proficiency as there is unchained access to the GPUs.

The main reason we choose to go with google colab is because it portable and can be accessed through anyone as long as they have a internet and a browser running. Here we can edit and run code at our ease without sitting together in one place.

Machine Learning

Machine Learning is the concept training a machine with patterns to predict the next outcome using the pattern. Here we are feeding the data to a machine and asking it to predict the future outcomes by analyzing the past outcomes through the experience it gained by being fed the past outcomes or datasets. There are many types of pre-defined machine learning algorithms that are already existing or we can create our own algorithms and teach it to the system.

Neural Networks

Neural networks are a sequence of operations that try to mimic the human brain to recognize the relationship between different types of data. They are inspired by the working of biological neural networks. It is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.

Today Neural Networks are used to solve many business problems.

7. IMPLEMENTATION

7.1 Data Collection

The data required for this work has been collected three different platforms. All three datasets are independent of each other, have different parameters and belong to different category of sales in different means but they all correspond to the sales and changes in market trend.

The first dataset is “Social network advertising cell” which has been collected from Kaggle and contains 401 data samples. This data was generated by Facebook API developers in 2017 and contains many categorical values. The dataset contains “gender” with binary value, “age” describing age of people with continuous values, “estimated salary” with continuous values describing the salary and “purchased” the dependent variable containing binary value whether the customer has purchased a product or not.

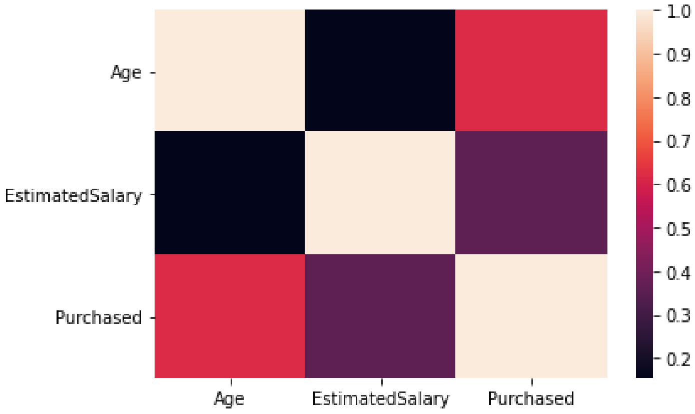
The second dataset is “Organic Purchased Indicator”. This dataset has also been obtained from Kaggle. The data samples were collected from a supermarket from January 2019 to the end of year relating to its customers and aimed at training a model to determine whether a customer would buy a product or not. The dataset consists of 13 different features with 22,000 data samples. The dataset contains the following features: “gender”, “geographic region”, “loyalty status”, “neighborhood cluster-55 level”, “neighborhood cluster-7 level”, “television region”, “affluence grade”, “age”, “frequency”, “frequency percent”, “loyalty card tenure”, “organics purchase count”, “organics purchase indicator” and “total amount spends”.

The third dataset is “Online Shoppers Purchasing Indicator” obtained from UCI Machine Learning repository. This dataset has been formed by google analytics for over a year. This data helps in determining whether a customer will purchase a product online in a website. This dataset contains 18 features and 12,330 datapoints. The dataset contains the following features: “Administrative”, “Administrative Duration”, “Informational”, “Informational Duration”, “Product Related”, “Product Related Duration”, “Bounce Rates”, “Exit Rates”, “Page Values”, “Special Day”, “Month”, “Operating Systems”, “Browser”, “Region”, “Traffic Type”, “Visitor Type”, “Weekend” and “Revenue”.

7.2 Data Analysis

Data analysis is an important step where we try to find the patterns in data. We consider three datasets individually to get insights from them. From the first dataset “Social Network Advertising Sells” we made the following analysis.

Figure 7.1: Heatmap for Correlation of features



From figure 1 we can see that Age is a feature that is strongly correlated to determining whether a person purchased a product and also it is observed that estimated salary is weakly correlated to determine the purchase of a product by customer.

Figure 7.2: Bar graph showing distribution of “Purchased” feature

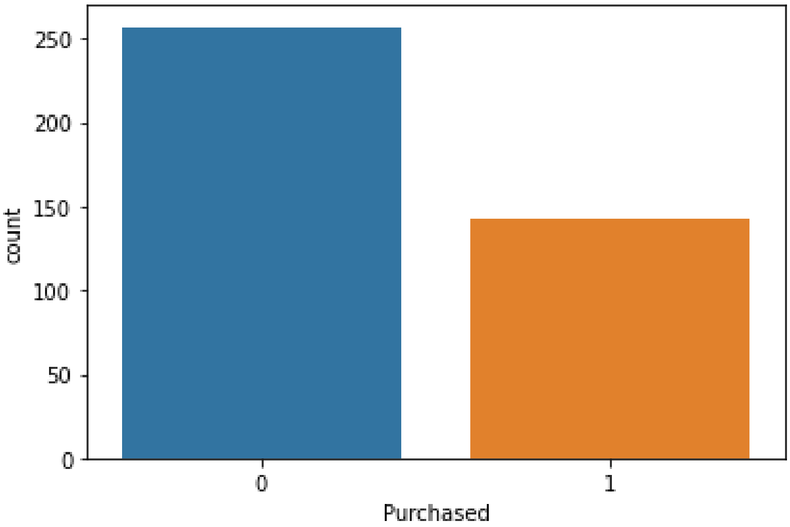
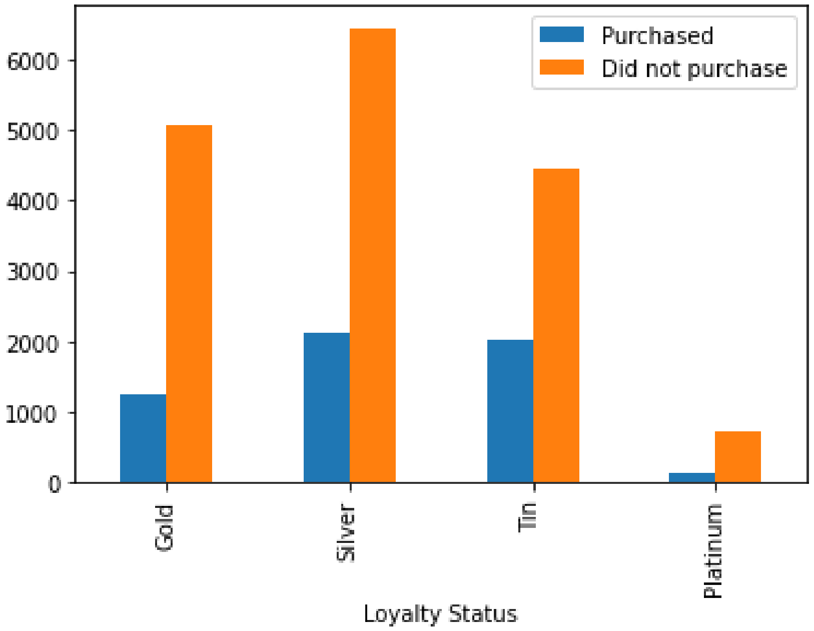


Figure 2 shows the distribution count of purchase of a product by customer. Around 60% records belong to customers who did not purchase the product and remaining 40% data is purchased data. This shows that the data is balanced.

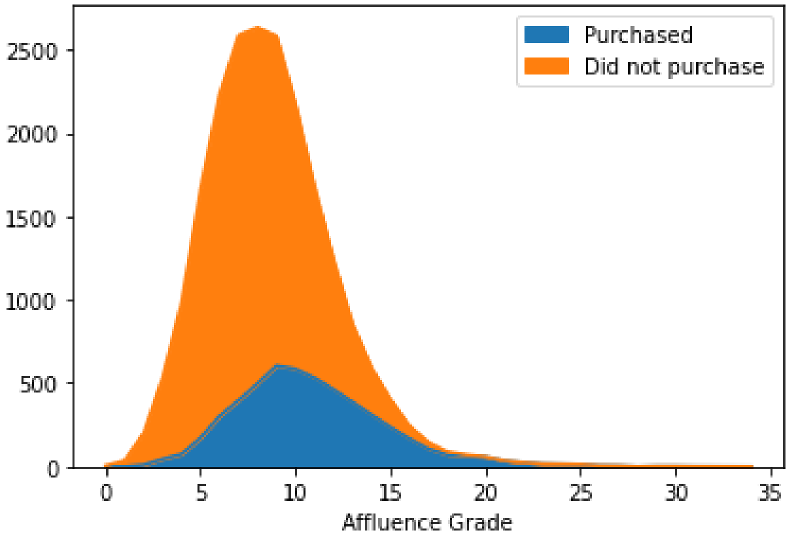
From the second dataset “Organic Purchase Indicator” we infer the following details.

Figure 7.3: Bar graph comparing Status with Purchase count



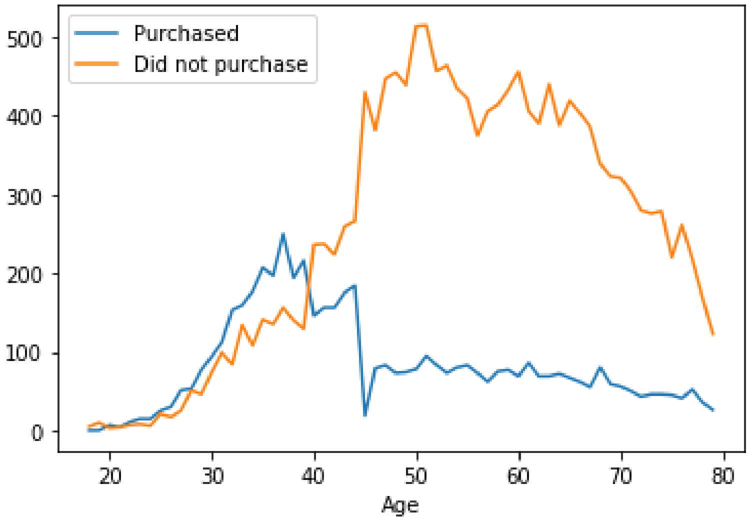
The customers are given loyalty status and it is observed that majority of the customers are in “silver” status, but majority of “tin” status customers purchase the products. Also, very less “platinum” members exist.

Figure 7.4: Comparing purchase count with Affluence grade



From figure 4 where affluence grade is mapped with purchase indicator, it is observed that affluence grade increase shows increase in purchases however the max point is reached at grade 10 from where as affluence grade increases the purchases decreases.

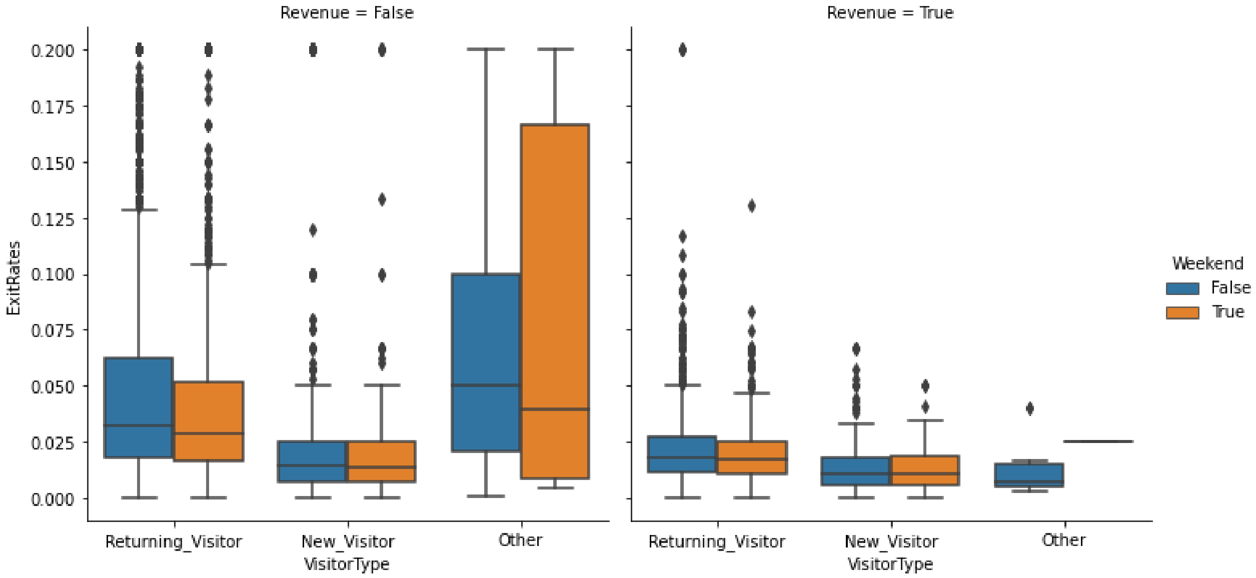
Figure 7.5: Line graph of Purchase count with respect to Age



From figure 5 it is observed that customers between ages 30 and 40 are observed to purchase more products than any other age groups.

From the third dataset “Online shoppers’ intention” it is observed that many customers irrespective of whether they have purchased visit the online store in weekends. When observed for purchases returning customers are people who purchase. This shows that the product is not going for a new customer and the market is revolving around the same old customers.

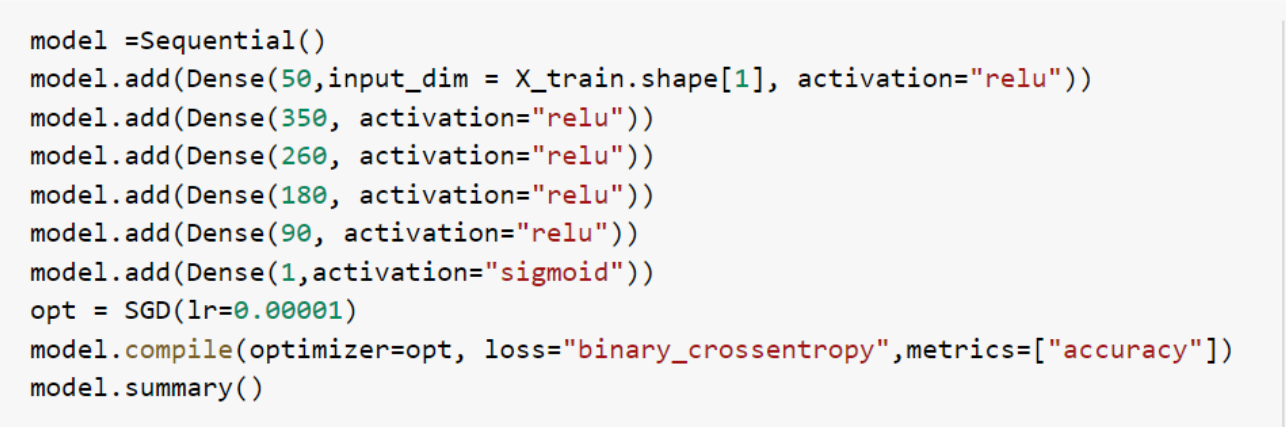
Figure 7.6: Box plot for Purchase count w.r.t New customer and Weekdays



7.3 Model Training

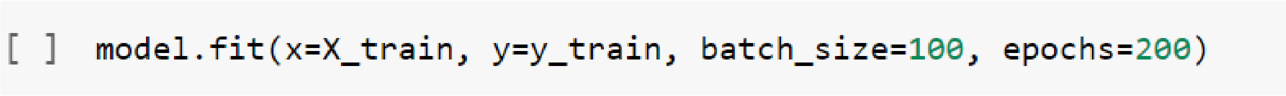
Initially a neural network is being constructed for all the three datasets. The artificial neural network contains an input layer of 50 nodes with 'relu' activation function. Input layer is followed by four hidden layers all having 'relu' activation function with 350, 260, 180 and 90 nodes, respectively. The output layer contains 1 node with sigmoid activation function. The loss function used is binary cross entropy as we must predict a binary class and the optimizer used is Stochastic gradient descent with a learning rate of 1e-5.

Figure 7.7: ANN model



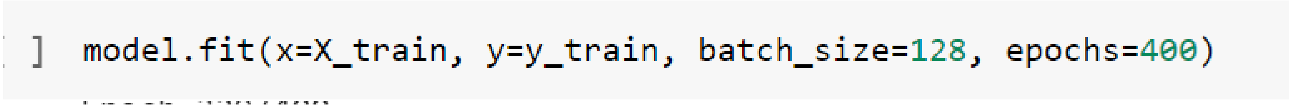
```
model =Sequential()
model.add(Dense(50,input_dim = X_train.shape[1], activation="relu"))
model.add(Dense(350, activation="relu"))
model.add(Dense(260, activation="relu"))
model.add(Dense(180, activation="relu"))
model.add(Dense(90, activation="relu"))
model.add(Dense(1,activation="sigmoid"))
opt = SGD(lr=0.00001)
model.compile(optimizer=opt, loss="binary_crossentropy",metrics=["accuracy"])
model.summary()
```

For the first data set, the run is for 200 epochs with batch size set to 100.



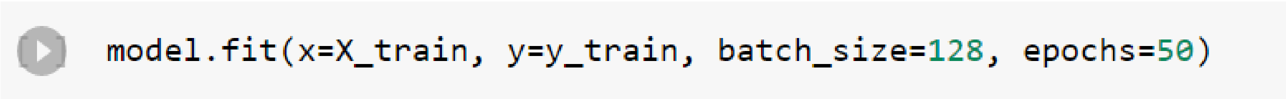
```
[ ]  model.fit(x=X_train, y=y_train, batch_size=100, epochs=200)
```

For the second data set, the run is for 400 epochs with batch size set to 128.



```
]  model.fit(x=X_train, y=y_train, batch_size=128, epochs=400)
```

For the final data set, the run is for 50 epochs with batch size set to 128.

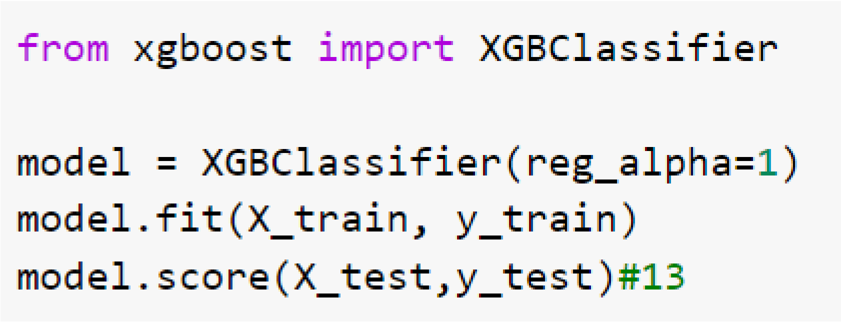


```
model.fit(x=X_train, y=y_train, batch_size=128, epochs=50)
```

The results obtained from the neural network were satisfactory, but the results had still scope of improvement. ‘Relu’ activation function is defined as max (0, x) where x in the input to the function.

Another model Extreme gradient booster is implemented to improve the results. Extreme Gradient Boosting Classifier uses the residual values calculated by using probabilities and actual values. These residual values are used to form a decision tree and are trained through it. The newly obtained values are regularized using learning rate and bias and again used to form a new tree. This process is repeated till the residual value approaches near zero values.

Figure 7.8: XGBoost model



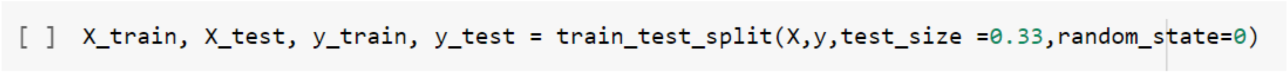
```
from xgboost import XGBClassifier

model = XGBClassifier(reg_alpha=1)
model.fit(X_train, y_train)
model.score(X_test,y_test)#13
```

7.4 Model Testing

The results obtained from the model are tested with the 33% test data that was split before training the model. The metrics used for testing is accuracy. Accuracy is the total number of correct predictions to the total values.

Figure 7.9: Model Testing

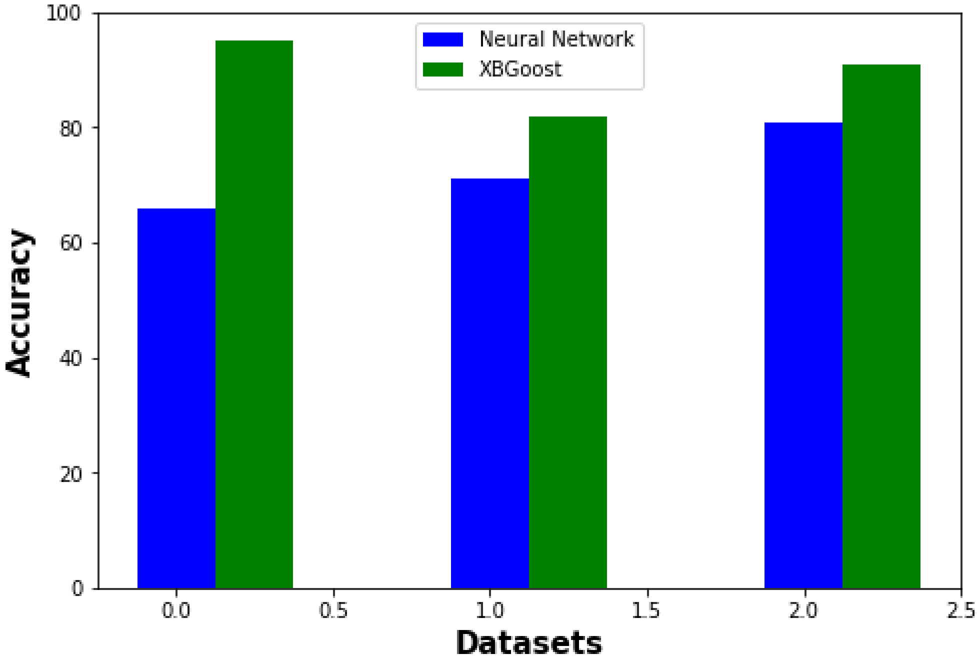


```
[ ]  X_train, X_test, y_train, y_test = train_test_split(X,y,test_size =0.33,random_state=0)
```

8. RESULTS AND DISCUSSIONS

The results obtained from the neural network are 66% for “Social network advertising cell” dataset, 71% for the “Organic Purchased Indicator” data set and 81% for “Online Shoppers Purchasing Indicator” data set. However, the results obtained for Extreme Gradient Boosting Classifier are improved. The results of Extreme Gradient Boosting Classifier for “Social network advertising cell” data set is 95%, for “Organic Purchased Indicator” data set is 82% and for “Online Shoppers Purchasing Indicator” data set is 91%.

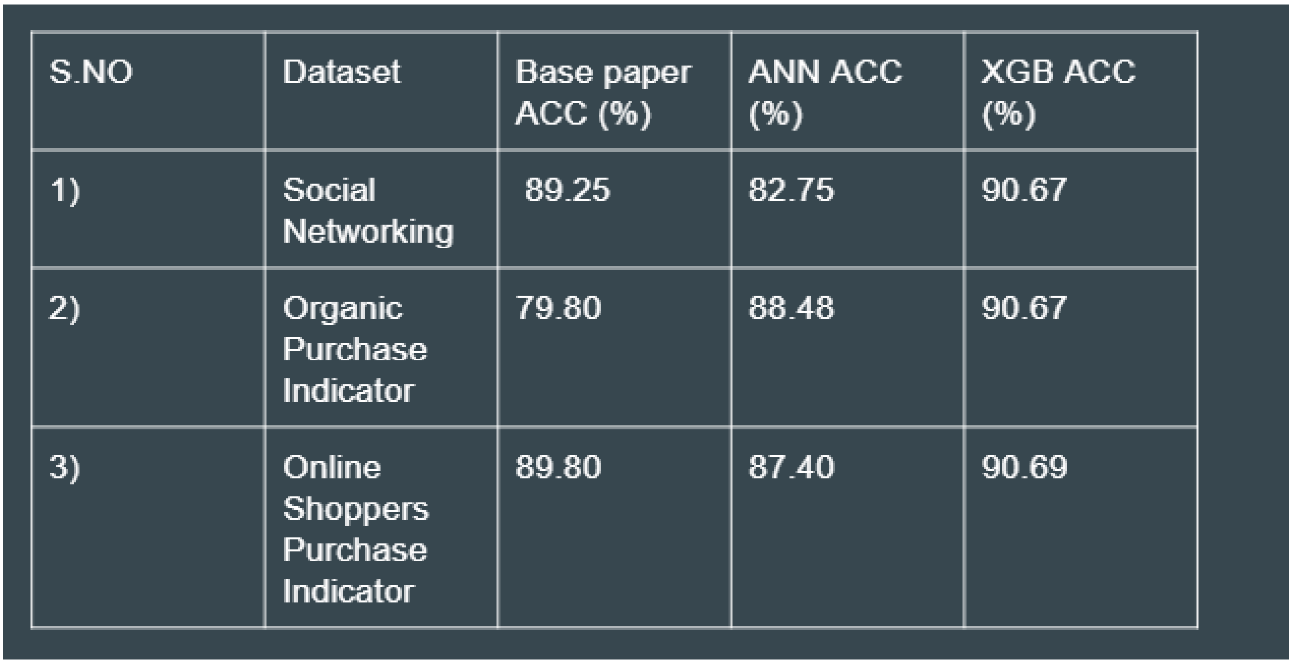
Figure 8.1: Bar graph comparing results.



As the figure above shows, the results obtained from the Extreme Gradient Boosting Classifier are better when compared to that of the artificial neural network constructed.

The result comparison in terms of numbers in the form of table is as follows:

Figure 8.2: Table comparing results.



S.NO	Dataset	Base paper ACC (%)	ANN ACC (%)	XGB ACC (%)
1)	Social Networking	89.25	82.75	90.67
2)	Organic Purchase Indicator	79.80	88.48	90.67
3)	Online Shoppers Purchase Indicator	89.80	87.40	90.69

9. CONCLUSION

In this work we have done an analysis on the data collected and have found parameters such as gender to be having high correlation towards the purchase of a product. This analysis can be used by business vendors to find the critical features that effect their business to reduce the risk of fluctuating market trends. Necessary precautions help business for a long stand also makes more profits than loss.

The results obtained from the Extreme Gradient Boosting Classifier algorithm were highest with the values 95%, 82% and 91% respectively for the three datasets. Artificial neural network constructed did not give the results as expected and might improve if more data samples were used to train the model.

The project can be further improved by adding more data samples. Also, as we have used three different data sets to cover a wide range of shopping types, a specific vendor can take data related to his business and get more accurate results. The data can be preprocessed for reducing the computational complexity even further and also new models can be applied to further improve the results.

10. REFERENCES

[1] Bloom, N. (2009). The Impact of Uncertainty Shocks. *Econometrica, vol 77, Issue 3*, 623-685.

[2] C. Vaca, D. R. (2020). Buy & Sell Trends Analysis Using Decision Trees. *2020 IEEE Colombian Conference on Applications of Computational Intelligence (IEEE ColCACI 2020)*, (pp. 1-6).

[3] Chen, T. a. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM.

[4] D. Soni, S. A. (2018). Optimised Prediction Model for Stock Market Trend Analysis. *2018 Eleventh International Conference on Contemporary Computing (IC3)*, (pp. 1-3).

[5] Jingye Lyu, M. C.-u.-d. (2020). Price volatility in the carbon market in China. *Journal of Cleaner Production*, 120-171.

[6] Ken, H. Z. (2013). Stocks market prediction using Support Vector Machine. *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering*, (pp. 115-118).

[7] M. Nabipour, P. N. (2020). Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis. *IEEE Access*, 150199-150212.

[8] Menon, A. a. (2019). A Review of Stock Market Prediction Using Neural Networks. *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, (pp. 1-6).

[9] Miró-Julià M., F.-R. G.-D. (2010). Decision Trees in Stock Market Analysis: Construction and Validation. *Trends in Applied Intelligent Systems* (pp. 185-194). Berlin, Heidelberg: Springer Berlin Heidelberg.

[10] Pelsser, A. G. (2020). Pricing and hedging in incomplete markets with model uncertainty. *European Journal of Operational Research*, 911-925.

[11] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, Elsevier BV*, 85–117.

[12] Usmani Mehak, E. M. (2018). Predicting Market Performance with Hybrid Model. *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)*, (pp. 1-4).