Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

Report On

Black Friday Sales Prediction

By

PA 13 Ritika Bhosale

PA 58 Sanyukta Tamhankar

PA 59 Gayatri Awate

PA 63 Vaishnavi Bhaitrat

Under the guidance of

Prof. Shilpa Sonawane

MIT-World Peace University (MIT-WPU)

Faculty of Engineering
School of Computer Engineering & Technology

*** 2021-2022***

MIT-World Peace University (MIT-WPU)

Faculty of Engineering
School of Computer Engineering & Technology

CERTIFICATE

This is to certify that Ritika Bhosale, Sanyukta Tamhankar, Gayatri Awate, Vaishnavi Bhaitrat of B.Tech., School of Computer Engineering & Technology, Trimester – IX /X, has successfully completed report on

Black Friday Sales Prediction

To my satisfaction and submitted the same during the academic year 2021-2022 towards the partial fulfillment of degree of Bachelor of Technology in School of Computer Engineering & Technology under Dr. Vishwanath Karad's MIT- World Peace University, Pune.

Prof. Shilpa Sonawane
Guide

Prof. Dr. Vrushali Kulkarni
Head
School of Computer Engineering & Technology

ACKNOWLEDGEMENT

I would like to express my gratitude towards our Guide Prof. Shilpa Sonawane for her invaluable guidance and support throughout the Seminar. Her comments and motivation helped a lot in completion of this project and report.

This really created a interest of deep learning of data science and developed the skill of problem solving in all of us.

Finally, I would like to thank School of Computer Engineering and Technology and MIT World Peace University for providing us with the platform which has led to many such work of good quality and value.

Thank- You

INDEX

Abstract

The largest shopping day of the year in America is the Friday following the Thanksgiving holiday. It is recognized as the ignition of one of the busiest shopping seasons in a year. From the computer science point of view, one of the most interesting applications of machine learning in the retail industry is to effectively predict how much a customer is probably to spend at a store based on historical purchasing patterns. If retailers comprehensively understand their customers in terms of characteristics, behaviors and motivations in the previous shopping seasons, they can implement and develop more effective marketing strategies for specific customers categories. This study proposes an empirical implementation of extreme gradient boosted algorithm for addressing an interesting challenge in the retail industry.

Keywords: retail industry, data science, intelligence, infrastructure, data mining

1.Introduction:

For a long history of several decades, Black Friday has been recognized as the largest shopping day of the year in the US. It is the Friday after Thanksgiving and for American consumers, it ignites the Christmas holiday shopping. For most retailers, it is the busiest day of the year. Black Friday is traditionally known for long lines of customers waiting
outdoors in cold weather before the open hours. Sales are so high for Black Friday that it has become a crucial day for stores and the economy in general with approximate 30% of all the annual retail sales occurring in the time from Black Friday through Christmas making it the kick-off day for the busiest and most profitable season for many businesses. It is unofficially a public holiday in more than 20 states and is considered the start of the US Christmas shopping season. In 2018, US shoppers expected to drop $483.18 on the shopping holiday of holidays, which equates to $90.14 billion. Although Black Friday is originally from America, it has become a universal recognition worldwide.

Because consumers are eager to spend so much money during this period, retailers seriously look forward to good preparation for the shopping holiday. In preparation for this day, retailers will typically hire more employees, stock their commodities, prepare new promotions, and decorate store layouts. Retailers rely on designing advertising
campaigns to attract more customers into their stores and/or their online shops. In order to maximize their efforts and revenues, retailers enthusiastically understand how the consumers make shopping decisions that will assist them to achieve the most profits during the shopping season. Many possible parameters that have been considered and are presented. If retailers comprehensively understand their customers in terms of characteristics, behaviors and motivations in the previous shopping seasons, they can implement and develop more effective marketing strategies for specific customers categories.

2. Motivation-

Black Friday sales in US still accounts for a whopping $6 Billion in revenue. In order to compete with Online Shopping Platforms, Brick and Mortar based Retailers need to figure out how to boost Sales during the most important Shopping Day of the Year. By understanding the Purchase Patterns of the Customers Retailers can provide improved Service Quality. Improve Staffing and Inventory of the Retail Store.

3. Problem Definition-

A sales forecast helps every business make better business decisions. It helps in overall business planning, budgeting, and risk management.

Sales forecasting allows companies to efficiently allocate resources for future growth and manage its cash flow.

Sales forecasts help sales teams achieve their goals by identifying early warning signals in their sales pipeline and course-correct before it's too late

Sales forecasting also helps businesses to estimate their costs and revenue accurately based on which they are able to predict their short-term and long-term performance.

4. Objectives:

- **Predicting Purchase -**

Build a simple Machine Learning model that can predict how much a customer is likely to spend on the eve of Black Friday.

- **Pattern Recognition -**

Reveal and understand the most important factors from predictors such as Age, Gender, City of Residence etc., that influence the spending of a customer. Establish a quantitative impact of the revealed factors and how they influence Purchase by a Customer on a personal level i.e., whether they have a positive or negative contribution on the Purchase.

5. Software Requirements:

1. Windows os
2. Jupter Notebook
3. Anaconda Navigator
4. Python 6.7

6. Data Preprocessing tasks performed:

Most of the raw data contained in any given Dataset is usually unprocessed, incomplete, and noisy. In order to be useful for data mining purposes, the Dataset needs to undergo pre-processing, in the form of ‘Data Cleaning’ and ‘Data Transformation’.

Handling Missing

Values Handling Outliers

Dealing with Categorical Variable

In our dataset the only predictors having missing value are Product_Category_1 and Product_Category_2. We can either try to impute the missing values or drop these predictors. We can text both approaches to see which returns the best results.

7.Exploratory Data Analysis (EDA):

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

In our project we had performed Univariate Analysis, Bivariate Analysis.

Univariate analysis:

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

Bivariate analysis:

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis.[1] It involves the analysis of two variables (often denoted as *X*, *Y*), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. Bivariate analysis can help determine to what extent it becomes easier to know and predict a value for one variable (possibly a dependent variable) if we know the value of the other variable (possibly the independent variable) (see also correlation and simple linear regression)

Collect patient data

Chatbots can extract patient information using simple questions about name, address, symptoms, current doctor, and insurance details. Chatbots then store this information in the medical facility system to facilitate patient admission, symptom tracking, doctor-patient communication, and medical record keeping

8. Model building using ML for key decisions:

1.Linear Regression:

Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple.

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other.

2. Random Forest:

Random forests or **random decision forests** are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.[1][2] Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.

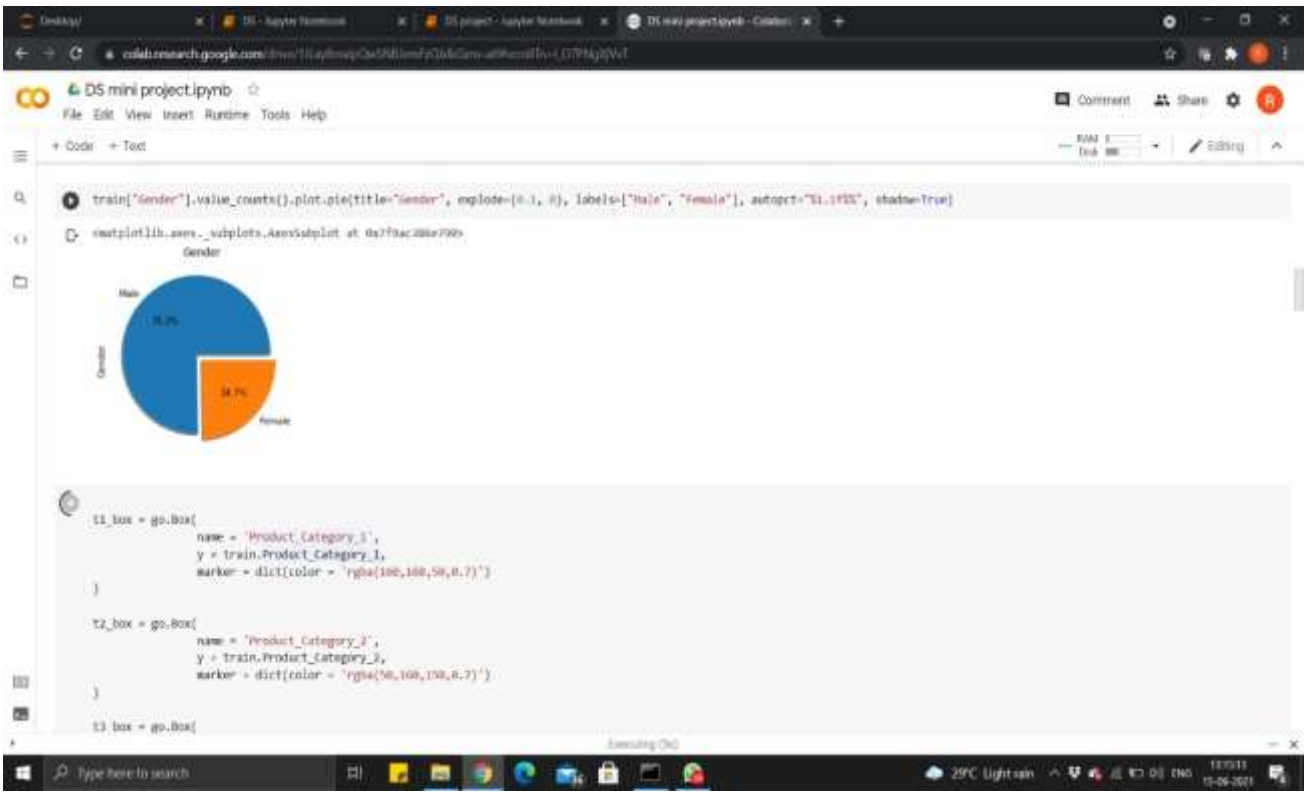
3. Decision Tree:

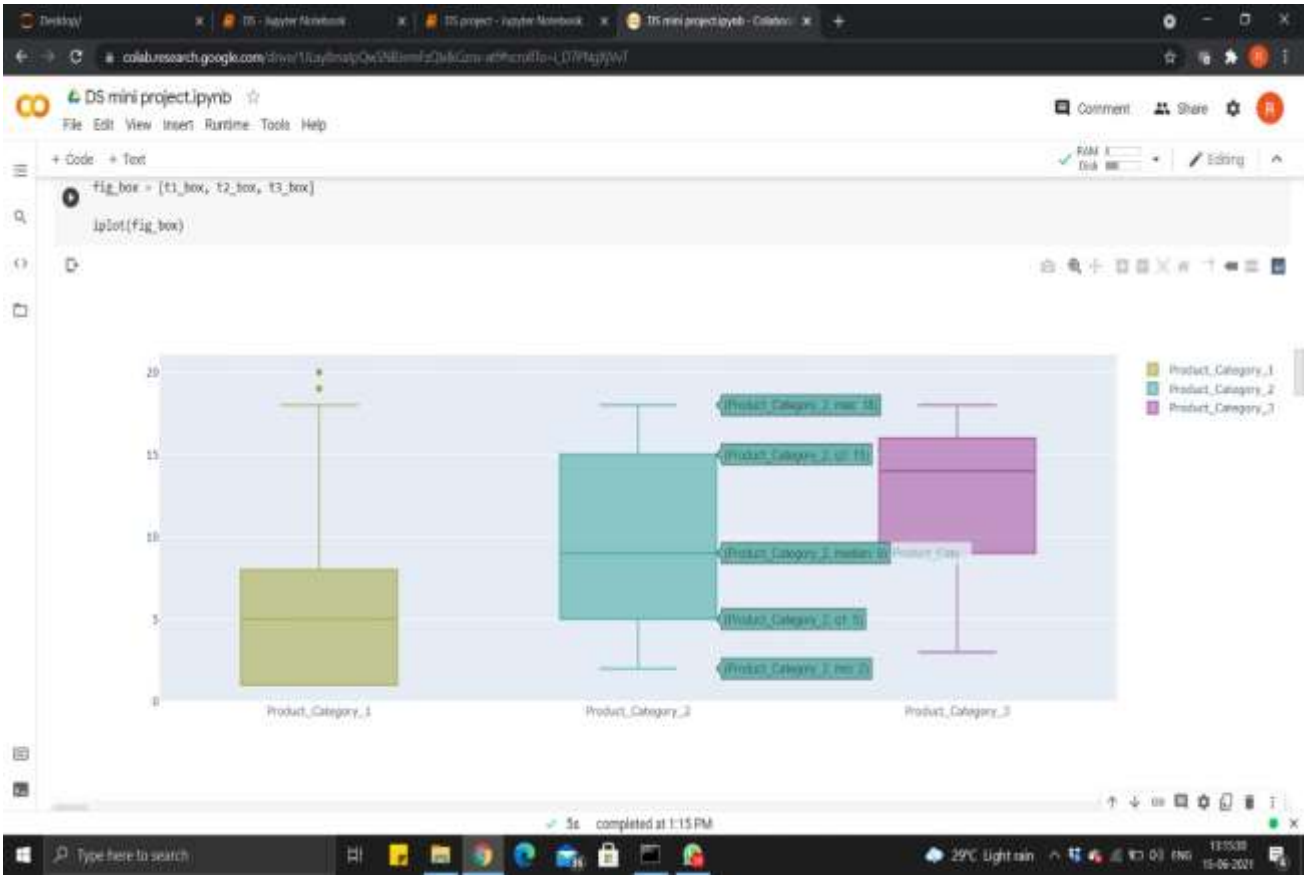
A Decision Tree is an algorithm used for supervised learning problems such as classification or regression. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature.

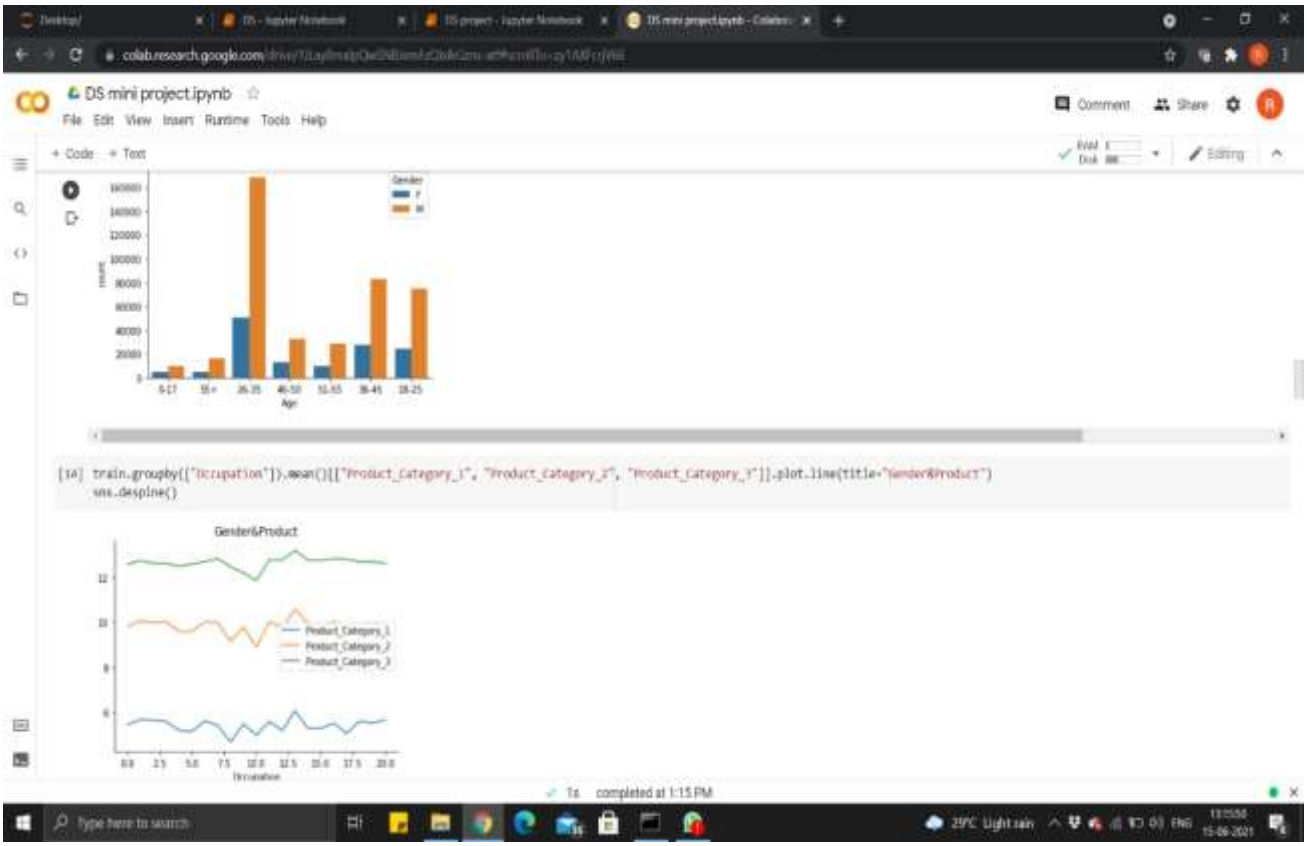
A **decision tree** is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

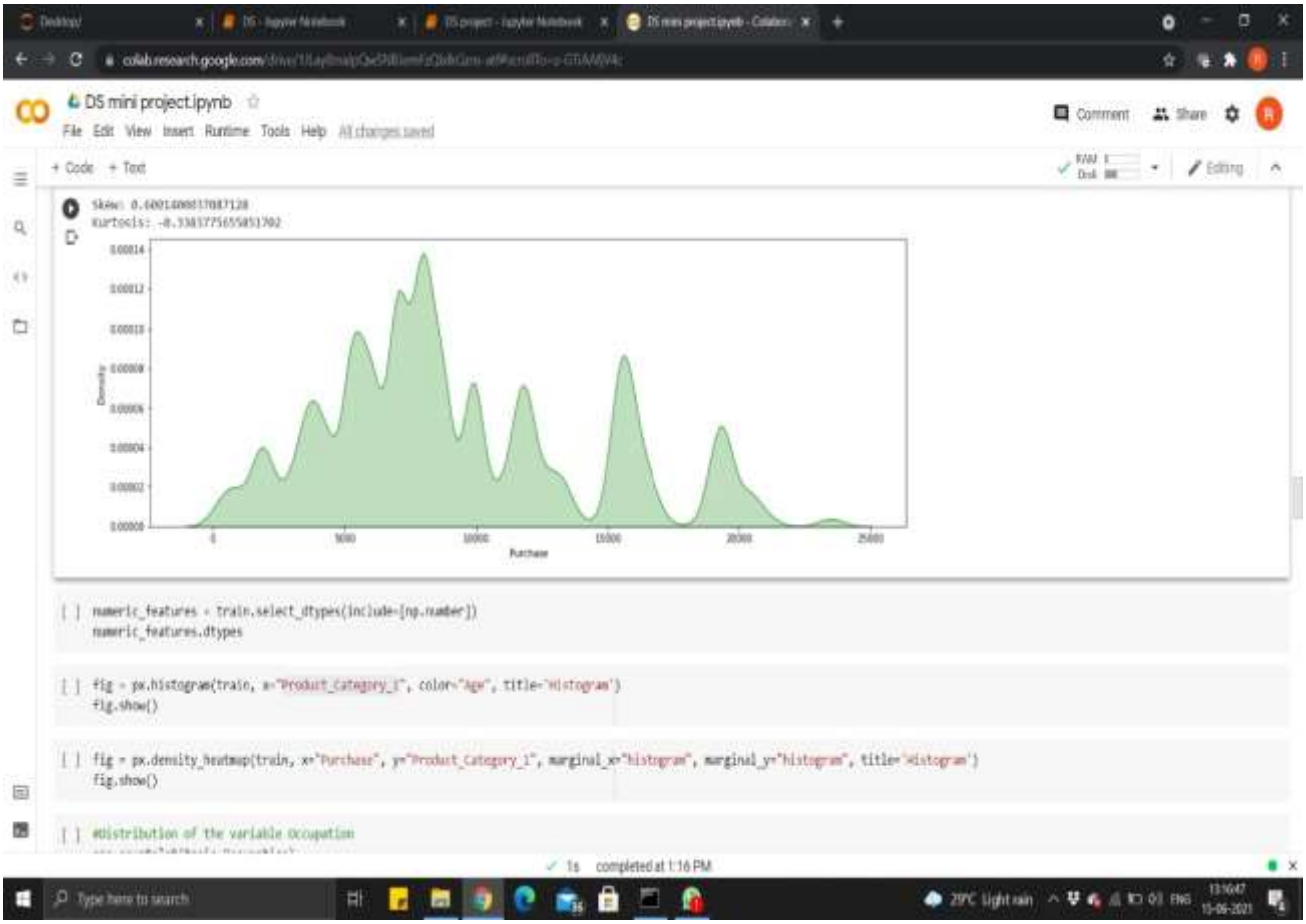
Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

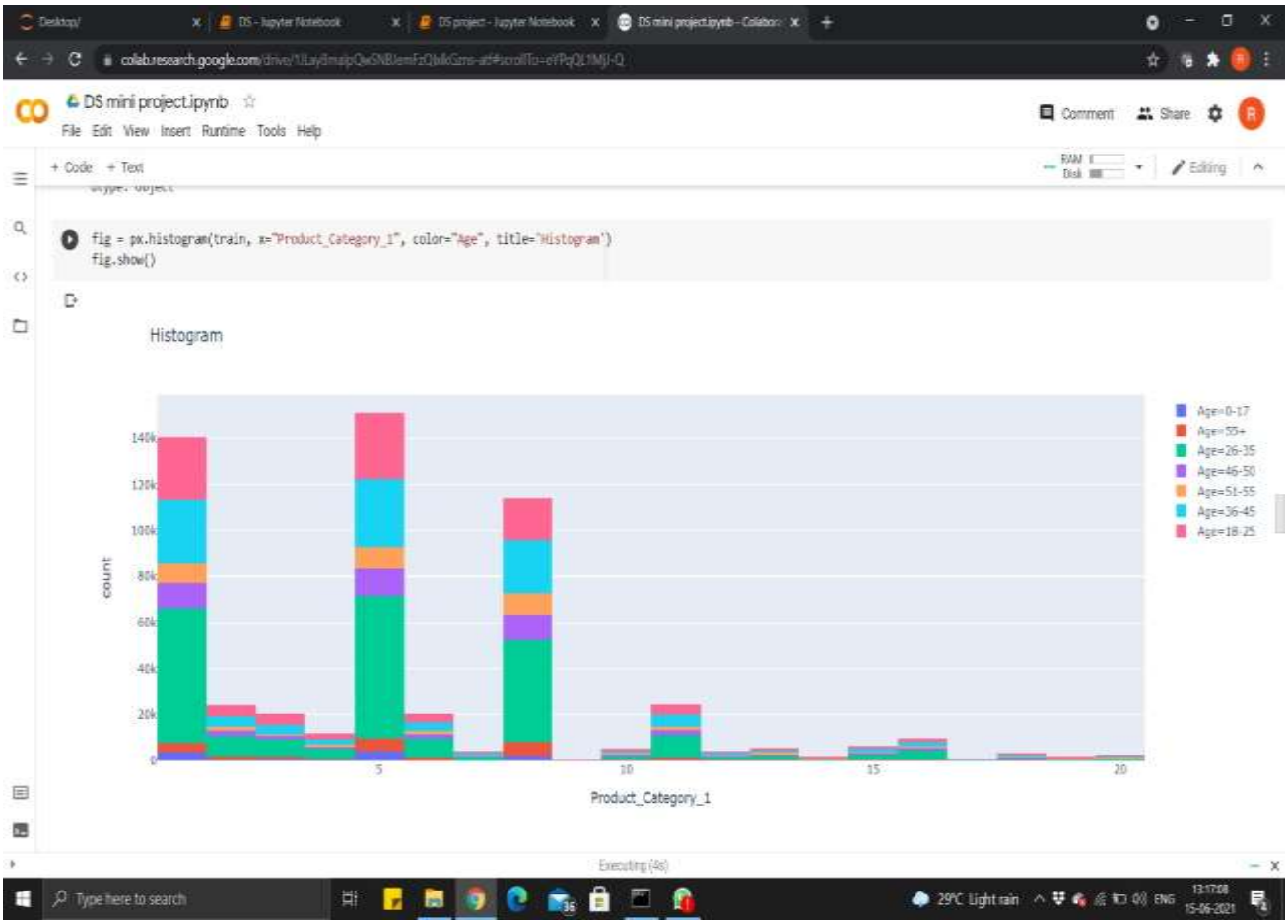
9.Screenshots of output:

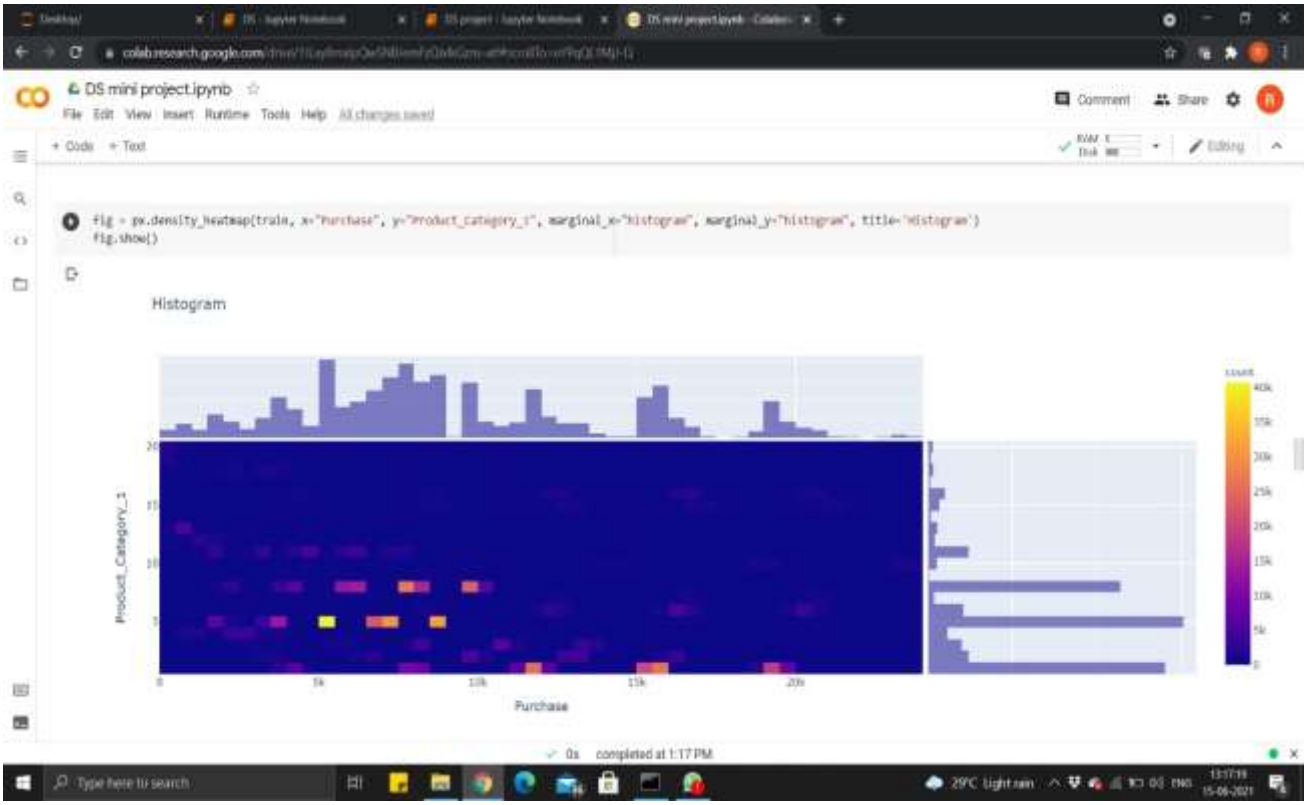


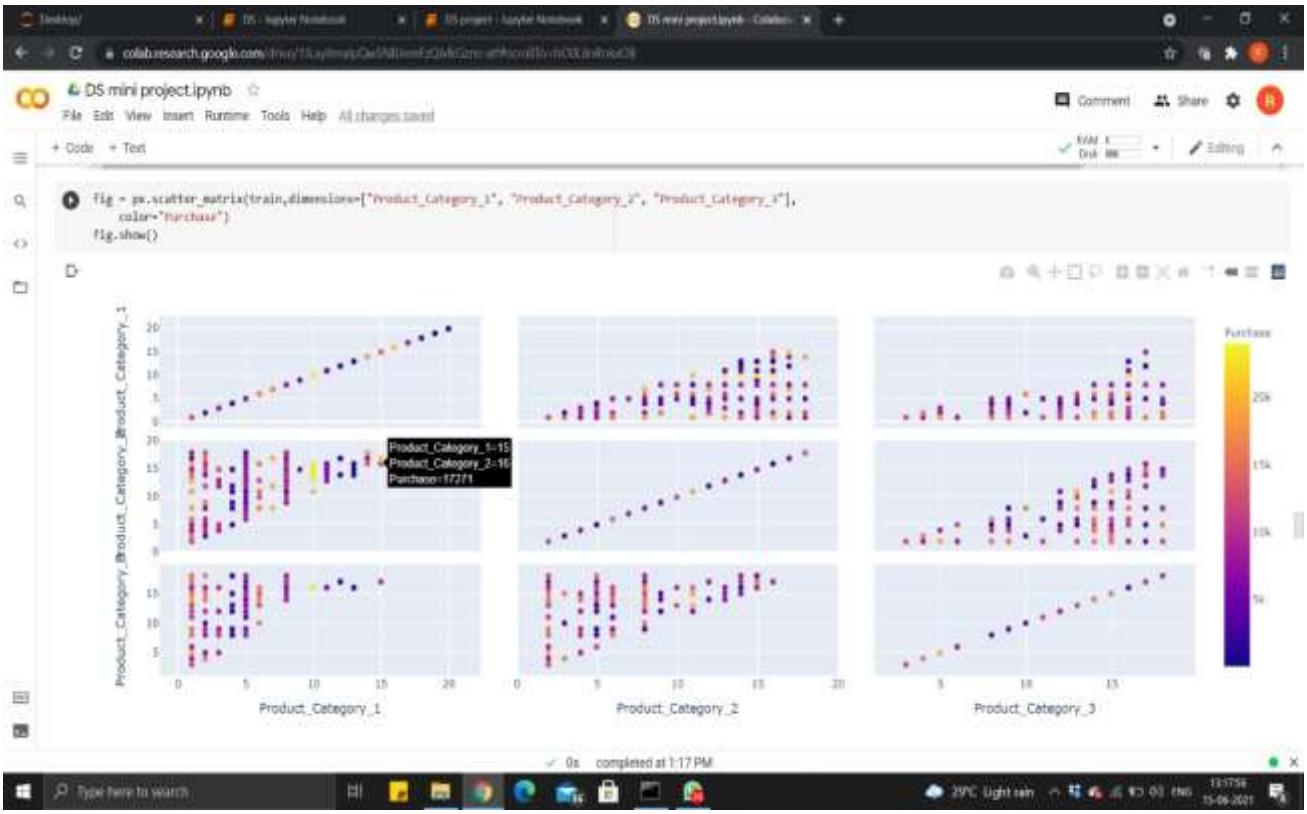


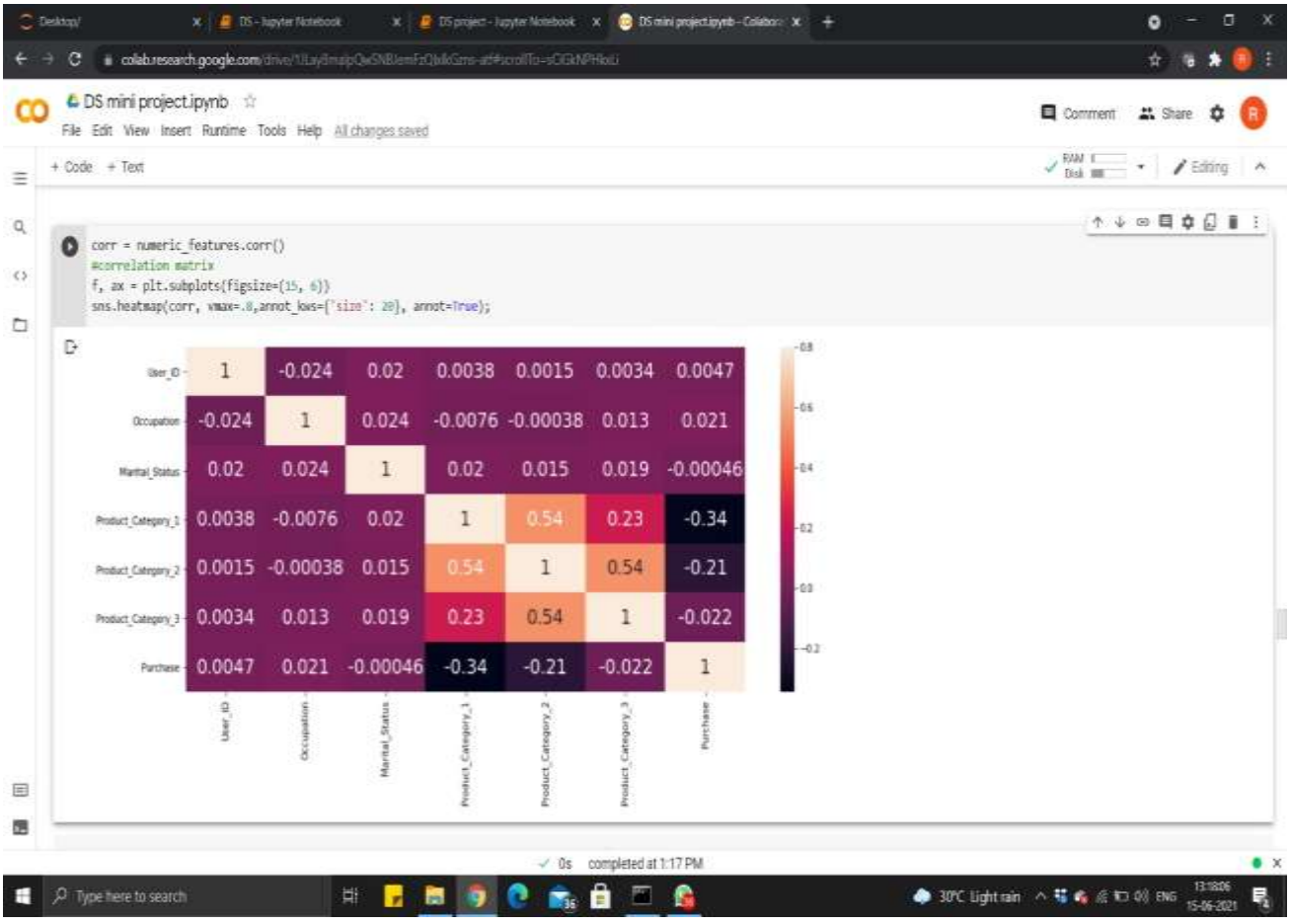

DS mini project.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
Comment
Share
Editing
Age
[14] train.groupby(["Occupation"]).mean()[["Product_Category_1", "Product_Category_2", "Product_Category_3"]].plot.line(title="Gender&Product")
sns.despine()
Gender&Product
Product_Category_1
Product_Category_2
Product_Category_3
1s completed at 1:15 PM
Type here to search
29°C Light rain

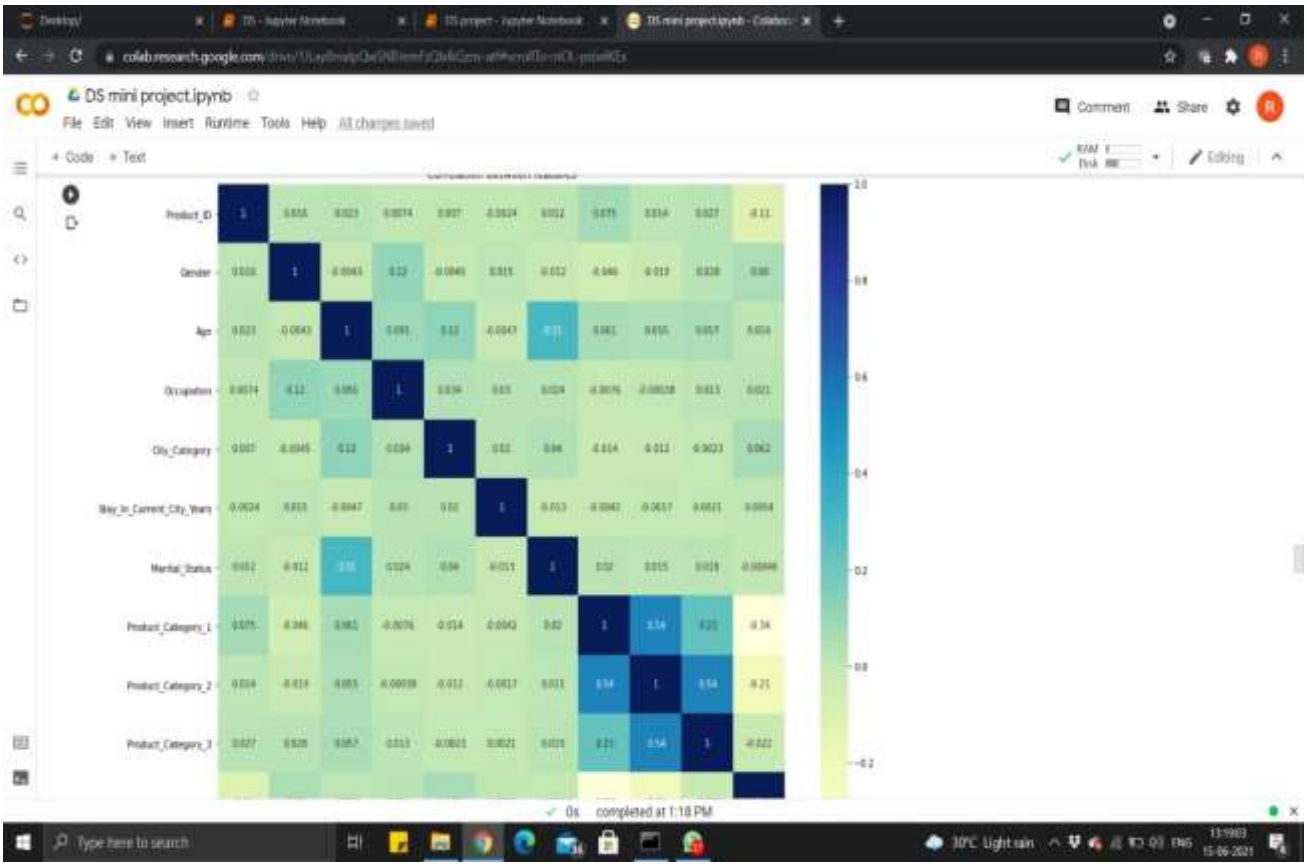

DS mini project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Comment
Share
Editing
Skew: 0.6001400017087128
Kurtosis: -0.33837756550851702
Density
Purchase
[ ] numeric_features = train.select_dtypes(include=[np.number])
numeric_features.dtypes
[ ] fig = px.histogram(train, x="Product_Category_1", color="Age", title='Histogram')
fig.show()
[ ] fig = px.density_heatmap(train, x="Purchase", y="Product_Category_1", marginal_x="histogram", marginal_y="histogram", title='Histogram')
fig.show()
[ ] #distribution of the variable Occupation
1s completed at 1:16 PM
Type here to search
29°C Light rain

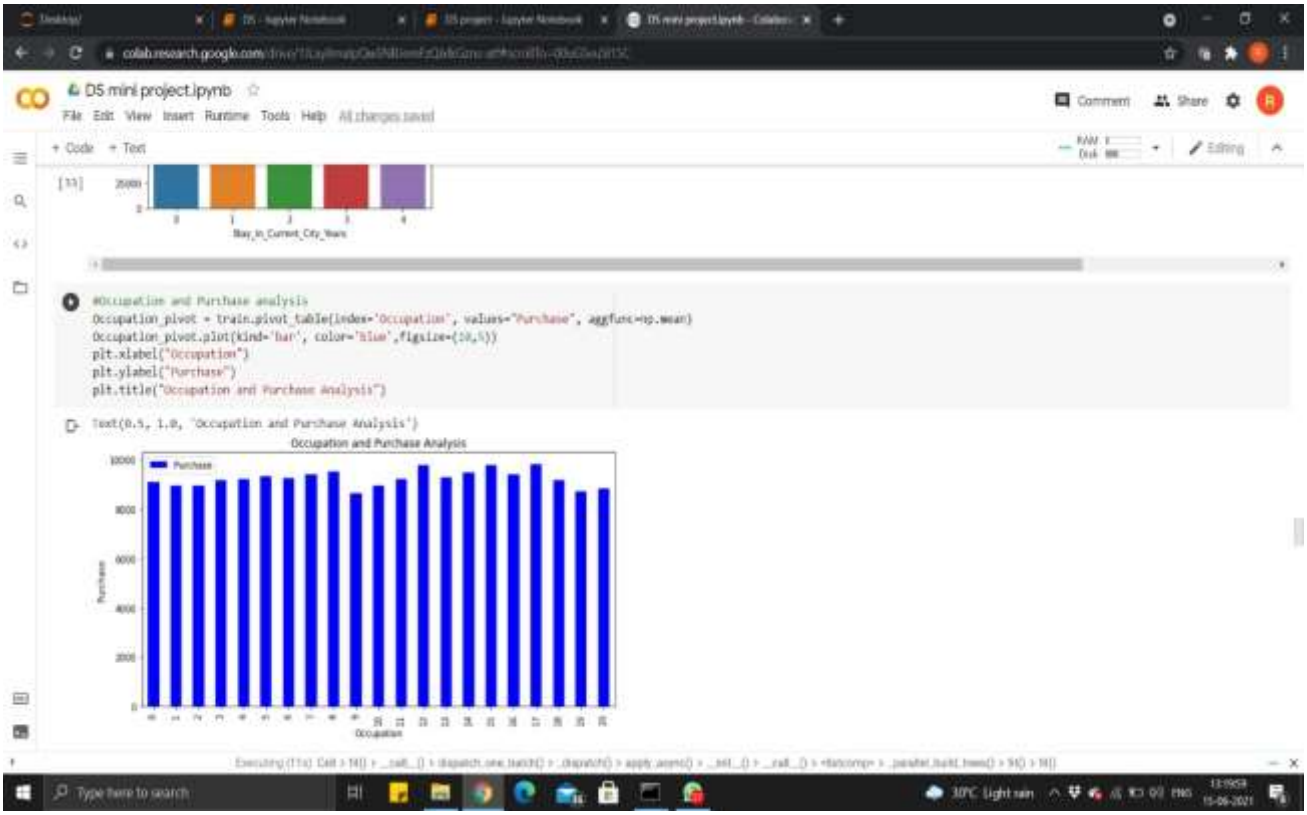

DS mini project.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
Editing
fig = px.histogram(train, x="Product_Category_1", color="Age", title='Histogram')
fig.show()
Histogram
count
Product_Category_1
Age=0-17
Age=55+
Age=26-35
Age=46-50
Age=51-55
Age=36-45
Age=18-25


DS mini project.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
Editing
fig = px.density_heatmap(train, x="Purchase", y="Product_Category_1", marginal_x="histogram", marginal_y="histogram", title='Histogram')
fig.show()
Histogram
Product_Category_1
Purchase


DS mini project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
fig = px.scatter_matrix(train,dimensions=["Product_Category_1", "Product_Category_2", "Product_Category_3"],
color="Purchase")
fig.show()
Product_Category_1
Product_Category_2
Product_Category_3
Purchase
0s completed at 1:17 PM


DS mini project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
corr = numeric_features.corr()
#correlation matrix
f, ax = plt.subplots(figsize=(15, 6))
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
0s completed at 1:17 PM


DS mini project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code
+ Text
Product_ID
Gender
Age
Occupation
City_Category
Stay_In_Current_City_Years
Marital_Status
Product_Category_1
Product_Category_2
Product_Category_3
completed at 1:18 PM
Type here to search
30°C Light rain


DS mini project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code
+ Text
Stay_In_Current_City_Years
#Occupation and Purchase analysis
Occupation_pivot = train.pivot_table(index='Occupation', values="Purchase", aggfunc=np.mean)
Occupation_pivot.plot(kind='bar', color='blue',figsize=(10,5))
plt.xlabel("Occupation")
plt.ylabel("Purchase")
plt.title("Occupation and Purchase Analysis")
Text(0.5, 1.0, 'Occupation and Purchase Analysis')
Occupation and Purchase Analysis
Purchase
Occupation
Type here to search
30°C Light rain

10.Advantages and Disadvantages:

Advantages:

- **Budgeting** – Companies can use predictive analytics to forecast more accurately their budgeting needs, instead of having to speculate and rely on the old models of "what-if". As a result, collaborations between departments can be improved for better team-works.
- Customers' insight– as mentioned above, advanced analytics can help businesses produce actionable customer insights predicting future actions by consumers. Companies can use this information to create better products/services tailored specifically to their customers. Similarly, companies can also apply this principle to increase the conversion rates as well as improve customers' loyalty, reward program, and more.
- Cost reduction - With the customer lifecycle becoming shorter and getting more complex, adopting predictive analytics and machine learning technology will help companies to have more effective marketing campaigns, resulting in the reduction of expenses while generating more revenue.
- Gaining perspective – Business organizations can implement predictive analytics to gain insights into the future success of their new products and/or services. This is especially helpful when there is insufficient historical data available to make a forecast or when the past is not indicative of the future.

Disadvantages:

- **Time-Intensive Completion** - While there are various methods of sales forecasting, the two broad approaches include manual and data-driven processes. In either case, significant time is required to develop forecasts. Within a traditional manual system, salespeople prepare their own forecasts by reviewing current accounts and overall projected sales. The time spent forecasting is less time spent selling. In more data-driven processes, a company often has marketing, IT and sales staff involved in building a system to collect and interpret data.
- **Expensive Technology Tools** - Manual processes are not as technology-oriented, but computer tools such as spreadsheets are commonly used. Typical sales organizations will also use database software to monitor ongoing relationships with customers. As you collect and analyze data in preparing forecasts, the greater its hardware and software program requirements. Companies will pay licensing fees to software providers for access. If each

salesperson has account access to use in managing relationships and preparing forecasts, the bill can get hefty for an organization.

- **Internal Bias -** Forecasting is not always intended to be a realistic projection of anticipated sales and not a depiction of desired sales. The challenge for company marketing and sales reps in preparing forecasts is that internal bias is hard to avoid

11.Conclusion:

All level participation will help companies internally assess their current business conditions, identifying the biggest weaknesses and opportunities for growth - in order to determine if predictive analytics can help to solve those business challenges and drive growth.

Sales forecasting plays a vital role in the business sector in every field. With the help of the sales forecasts, sales revenue analysis will help to get the details needed to estimate both the revenue and the income. Different types of Machine Learning techniques such as Support Vector Regression, Gradient Boosting Regression, Simple Linear Regression, and Random Forest Regression have been evaluated on food sales data to find the critical factors that influence sales to provide a solution for forecasting sales. After performing metrics such as accuracy, mean absolute error, and max error, the Random Forest Regression is found to be the appropriate algorithm according to the collected data and thus fulfilling the aim of this thesis

12.References

[1] Yang C, Li C, Wang Q, Chung D, Zhao H. Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. *Front Genet.* 2015;6:229. [PMC free article] [PubMed] [Google Scholar]

[2] Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform.* 2015;19:1209–15. [PubMed] [Google Scholar]

[3] Kankanhalli A, Hahn J, Tan S, Gao G. Big data and analytics in healthcare: introduction to the special section. *Inform Syst Front.* 2016;18:233–5. [Google Scholar]

[4] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inform Sci Syst.* 2014;2:3. [PMC free article] [PubMed] [Google Scholar]

[5] Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. –Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Trans Biomed Eng.* 2017;64:263–73. [PMC free article] [PubMed] [Google Scholar]

[6] Wang Y, Kung LA, Wang WY, Cegielski CG. An integrated big data analytics-enabled transformation model: application to health care. *Inf Manag.* 2017;55:64–79. [Google Scholar]

[7] El-Gayar O, Timsina P. Opportunities for business intelligence and big data analytics in evidence based medicine. System Sciences (HICSS); 47th Hawaii international conference on 2014.2014. pp. 749–57. [Google Scholar]

[8] Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *Int J Med Inform.* 2017;98:22–32. [PubMed] [Google Scholar]

[9] Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. *Proteomics.* 2016;16:741–58. [PubMed] [Google Scholar]

[10] Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. *Biomed Inform Insights.* 2016;8:1. [PMC free article] [PubMed] [Google Scholar]

[11] Image of medical records. Osadcha, Juliia. Web UI Color [.png]. Retrieved from URL https://www.iconfinder.com/Juliia_Os

[12] Image of prescription, The Recycling Partnership. Recycling Extras [.png]. Retrieved from URL https://www.iconfinder.com/TheRecyclingPartnership

[13] Osadcha, Juliia. Web UI Color [.png]. Retrieved from URL https://www.iconfinder.com/Juliia_Os

[14] Image of patient. Babic, Goran. Creative Agency Bresign. Medical[.png] Retrieved from URL https://www.iconfinder.com/Bres

[15] Image of analytic representation. Retrieved from URL

https://store.kde.org/usermanager/search.php?username=S ephiroth6779

[16] Image of hospital sign. Babic, Goran. Creative Agency Bresign. Medical[.png] Retrieved from URL https://www.iconfinder.com/Bres

[17]Access to Health Services. (2015, October) Retrieved from URL http://www.healthypeople.gov/2020/topicsobjectives/topic/Access-to-Health-Services

[18] World Health Organization Statistical Profile. (2015), Retrieved from URL's: http://www.who.int/gho/countries/chn.pdf?ua=1, http://www.who.int/gho/countries/usa.pdf?ua=1

[19] Agarwal M, Adhil M, Talukder AK. *International Conference on Big Data Analytics.* Cham, Switzerland: Springer International Publishing; 2015. Multi-omics multi-scale big data analytics for cancer genomics; pp. 228–43. [Google Scholar]

[20] He KY, Ge D, He MM. Big data analytics for genomic medicine. *Int J Mol Sci.* 2017;18:412. [PMC free article] [PubMed] [Google Scholar]

[21] Tan SL, Gao G, Koch S. Big data and analytics in healthcare. *Methods Inf Med.* 2015;54:546–7. [PubMed] [Google Scholar]

[22] Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M. et al. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS One.* 2016;11:e0157077. [PMC free article] [PubMed] [Google Scholar]

[23] Costa FF. Big data in biomedicine. *Drug Discov Today.* 2014;19:433–40. [PubMed] [Google Scholar]

[24] Yao Q, Tian Y, Li PF, Tian LL, Qian YM, Li JS. Design and development of a medical big data processing system based on Hadoop. *J Med Syst.* 2015;39:23. [PubMed] [Google Scholar]

[25] Kambatla K, Kollias G, Kumar V, Grama A. Trends in big data analytics. *J Parallel Distrib Comput.* 2014;74:2561–73. [Google Scholar]