DATAXPLORER USING DATA SCIENCE

Major project report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology

in

Computer Science and Engineering

Submitted by:

Astha Kapil (181242)

UNDER THE SUPERVISION OF

Prof. Dr. Vivek Kumar Sehgal Head Of Deptt. Of CSE & IT

Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Waknaghat, Solan - 173234,
Himachal Pradesh, INDIA

DECLARATION

I hereby declare that this project has been done by me under the supervision of Prof. Dr. Vivek Kumar Sehgal, Head Of Department Of CSE & IT, Jaypee University Of Information Technology. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:
Prof. Dr. Vivek Kumar Sehgal
Head Of Department Of CSE & IT
Department of Computer Science & Engineering and Information Technology
Jaypee University Of Information Technology

Submitted by:
Astha Kapil (181242)
Computer Science & Engineering Department
Jaypee University Of Information Technology

CERTIFICATE

I hereby declare that the work presented in this report entitled **" Dataxplorer Using Data Science"** in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from January 2022 to May 2022 under the supervision of **Prof. Dr. Vivek Kumar Sehgal** Head Of Department Of CSE & IT.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Astha Kapil (181242)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Prof. Dr. Vivek Kumar Sehgal
Head Of Department
CSE & IT

ACKNOWLEDGEMENT

I take the opportunity to express my heartiest thanks and gratefulness to almighty God for their divine blessing makes it possible to complete the project work successfully.

I am really grateful and wish my profound indebtedness to my project supervisor Prof. Dr.Vivek Sehgal, Head Of Dept.Of CSE & IT, JUIT, Waknaghat for his deep Knowledge & keen interest in the field of Data Science to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project. I would also like to generously thank each one of those individuals who have helped me directly or indirectly in making this project a success. Finally, I must acknowledge with due respect the constant support and patience of my parents.

Astha Kapil (181242)
Computer Science & Engineering Department
Jaypee University of Information Technology

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Py	Python
Np	Numpy
Pd	Pandas
Info	Information
RFM	Recency, Frequency, Monetary
LTV	Lifetime value
B2C	Business to consumer
B2B	Business to business
ML	Machine learning
EDA	Exploratory data analysis
GUI	Graphical user interface
IOT	Internet of things
Std	Standard deviation
Corr	Correlation
Stat	Statistics
AOV	Average order value
AI	Artificial Intelligence

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Each and every company has their vision and to achieve that particular vision they need to keep some metrics in their mind and fulfill them to achieve it. Which basically includes “Determining their active consumer base”, "Finding the customer's lifetime value," "Predicting each customer's next purchase day," "Analysis the marketing campaigns," Depending on the type of firm/business and the services it mostly provides, and so forth. I visualize client behavior and characteristics from several perspectives in this project. I take it a step further and create a business case for whether or not the client base can be split to generate customized partnerships. I'll approach this topic from a behavioral standpoint to better understand client spending and ordering habits.(alternatives include regional and demographic approaches) also and marketing analysis. Basically it all will be a kind of data driven growth (using data for the growth of the company by analyzing and visualizing it). I utilize Python in a simple way to increase any company's growth by applying a predictive approach to all the actions in this project. The project will primarily use programming, data analysis, and machine learning. Starting with lifetime customer value, anticipating next purchase day, marketing analysis, and lastly understanding, analyzing, and visualizing the data, the entire process would be like this.

CHAPTER 1: INTRODUCTION

1.1 Introduction

Data science is a field in which we take raw data and apply technology to extract useful and relevant information. Data science necessitates a wide range of abilities, including computer science, statistics, ML and many more. These days, data science has a high demand in all the brands, startups, hospitals etc as they are using this to sort their data, to manage it and take out all the meaningful and important info out of it. Some of the most well-known applications of data science include recommendations, healthcare, targeted advertising, improved picture recognition, and much more.

In this project basically I visualized and analyzed the data, basically as we know that there are so many metrics/features on which the growth of a company or a startup depends. So in this project I tried to cover some of the best and most important factors which affect the growth of the company a lot. And once we get to know what is the right and wrong choice during decision making, it will definitely lead to improving the company's condition and growth solely. Last semester I started this project in which the covered factors were like finding the metric of the company, segmentation of the customer, churn prediction etc. Now in this semester I continue the project and focus on some of the most important features for any company that is finding the lifetime value of the customers/user which means to know if the customer will be a long term customer or a short term customer of the company, prediction of the next date when the consumer will buy, analyzation of marketing or advertising factors of the company like is there any positive results are coming by any advertising campaign or not or how impactful the campaign really is. Primarily I’m doing data driven growth, which means using data I’m checking and analyzing the growth (factors) of the company. A "data-driven" strategy means that a company bases its effective information research and interpreting considerations. Startups can use a data-driven process thoroughly and manage it accordingly to investigate the content or information they are having in bulk, for a better customer experience.

Some of the important terms used in this project are as follows:

RFM (Recency, frequency, monetary):
Recency: To compute this, we must first examine each consumer's most recent buy date and examine the days of their inactivity (that is how many it is). After we get the number of inactive days for each customer, we will use K-means clustering to assign them a recency score.

Frequency: In order to generate frequency clusters, you must first compute the overall amount of orders for every client.

Monetary:This is also referred to as revenue. As it implies, a customer earns or contributes a certain amount of revenue to the organization.

And all three depend on each other directly and indirectly. And I studied them briefly with the help of calculations and the visualizations in this project.

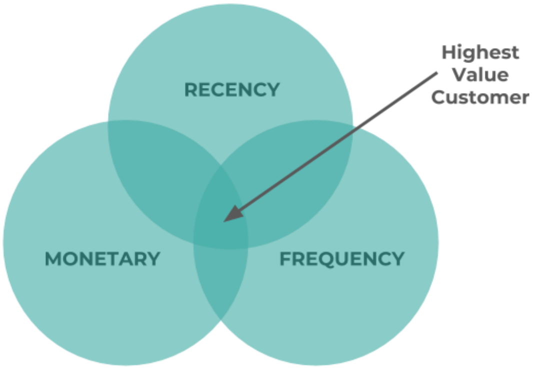


Fig: RFM model

In theory, for RFM, I'll have segments like these:

Low Value: Customers who are less dynamic than others, not extremely continuous purchaser/guest and produce exceptionally low - zero - possibly bad income.
Mid Value: Regularly utilizing our foundation (however not however much our High Values), genuinely continuous and produces moderate income.
High Value: The gathering we would rather not lose. High Revenue, Frequency and Low Inactivity.

Therefore, we can say that the high value means the high RFM value and low value indicates the lower RFM value.

Know your metrics: Each company should know their metric. The north star metric, is a metric/measurement which affects the growth of the company in the long run. So one should choose their metric wisely and it depends on the company's product, positions and targets etc.
For example: for facebook north star metric will the daily users, for spotify it is the time spent listening so it varies company to company.

Here in this project I'll be using the online retail dataset and in for this dataset, taking Revenue as the metrics as it is a dataset of online retail of xyz company. Dataset includes the majorly important columns such as Customers id, unit price, invoice date etc.
And for marketing analysis I'll be using a different dataset than this, that dataset includes columns that tell the customers personal information, columns showing the personal behavior of the customers and the responses of the customers based on the different columns. Having this dataset I can properly visualize and analyze the data and come up with the best campaign and best campaign techniques itself.

So, these are some of the terms that will be utilized in this project.

1.2 Problem Statement

Let's pretend there's a startup company in the corporate world. The first step in starting a new startup or the business is figuring out what problem one needs to solve. Which essentially means you must determine the source of the problem you will be solving. Which of your customers should be the focus one? What business are you focusing on? Is the aim business-to-business (B2B) or business-to-consumer (B2C)? Are you attempting to create a product? Or are you merely attempting to provide consultancy or technical assistance? Alternatively, delivering the services? Here’s some most crucial questions to consider when starting a company or expanding a current one so that you’ve a clear vision for your startup/company. In reality, the goal of this project is to assist startups, or any organization for that matter, in gaining insights into their consumers, and predicting how your firm will perform in the next 5 to 10 years. This will improve the sales and the revenue, which will determine the company's success. The impact your company has on the economy is much more important than the increase in revenue or money. I used the Analysis to look at a sample dataset and see how the company is assisting its clients. Is it having a positive or bad effect on them? Is it assisting them in making a positive difference in their daily lives? Are they able to retain them? Are they opting for the right campaigns? This is my problem statement: I'll use in-depth study of all conceivable elements to assist you determine whether your problem statement is being solved or not, and how to boost your company's growth.

1.3 Objective

This project's goal is to aid in the analysis and investigation of ways to boost a company's growth. and how a startup/business is performing in terms of crucial metrics (and how they can boost it also). As a result, I thought about a lot of things in order to examine them and try to come up with the best possible models and findings.

As a result, the primary purpose here is to properly examine all of the data in order to establish what is important. or useful to the firm and what isn't, and then act exclusively on that knowledge to enhance the company's growth and identify where your efforts should be focused. and where you should avoid wasting your energy and resources. Many times, firms devote all of their resources and money in the incorrect way for an extended length of time before realizing that it isn't working for them and that they need to modify their strategy to achieve the best results or profits. So, with all of that in mind, we should proceed.

This project, in which i go from predicting the lifetime value of a customer to prediction of the next purchase day of each customers to analyzing the marketing strategies in thoroughly to building the best possible model to get the best result out of it, will really help a company choose the right direction and come up with the most advantageous ways or strategies for more pay back.. What we're doing is data-driven growth, which means we took a dataset and ran a deep analysis on all of the variables to come up with the greatest possible solution for enhancing a company's growth, which is the main goal of this large enterprise.

All of this was accomplished with the help of data analysis, machine learning algorithms (such as linear regression, logistic regression, and k-means clustering), the elbow method, Jupyter Notebook, Python programming language, and other libraries imported, which will be discussed in the following pages.

1.4 Methodology

The first and most important step is to find a trustworthy source of pertinent information. The next stage is to assess the data and look for limits and conflicts after it has been loaded into the project. I moved on to data cleansing and preprocessing after I had a firm grasp on the properties. Following the necessary changes, the dataset must be optimized to achieve the desired results. Multiple machine learning methods, such as decision trees, various regressors, and classifiers, will be implemented on the dataset using various Python modules to develop future models and test their accuracy. I compared the ratings and analyzed the outcomes after receiving the data.

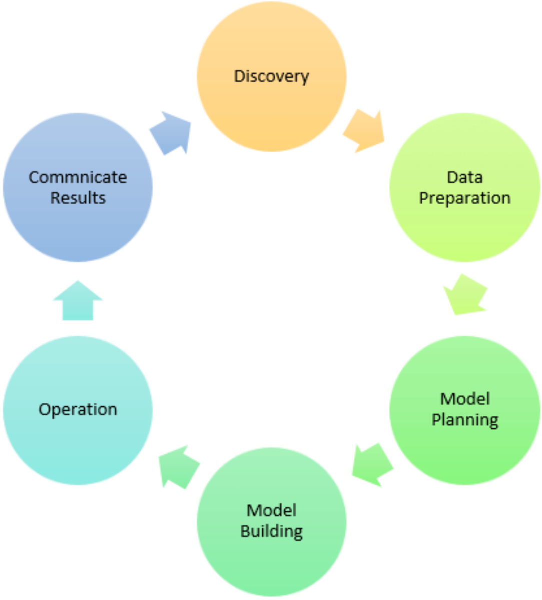


Fig: Data science processes

1. Discovery:

To answer the business question, the discovery process comprises obtaining information from all known internal and external sources.

The data could include:

A. Log files from the web server
B. Information obtained via social media
C. To stream data from web sources, census API datasets are employed.

2. Data Preparation:

Data discrepancies such as missing integers, blank columns, and an incorrect data format must all be cleansed. You must first process, investigate, and condition the data before you can model it. If your data is clean, your projections will be better.

3. Model Planning:

At this stage, decide on the method and strategy for drawing the link between input variables. Several statistical methodologies and visualization tools are used to plan the model. SQL analytic services, R, and others are examples.

4. Model Construction:

Now comes the fun part: constructing the prototype. In this part, data engineers provide samples for analysis and retraining. Techniques like association, categorization, or clustering are used in the training sample. After it has been built, every approach is verified even against "test" samples.

5. Actualization:

You send the final timescales system, including documentation, coding, any research papers, at this stage. The model is implemented in a real-time production environment after intensive testing.

6. Make Data Simple to Find:

At this time, all participants are informed of the principal results. One can assess if the project was a good or bad decision depending just on the model's inputs.

CHAPTER 2: LITERATURE SURVEY

2.1 Related Literature

S.NO.	Literature	Topic discussed
1.	S. Wu, W. -C. Yau, T. -S. Ong and S. -C. Chong, "Integrated Churn Prediction and Customer Segmentation Framework for Telco Business," in IEEE Access, vol. 9, pp. 62118-62136, 2021, doi: 10.1109/ACCESS.2021.3073776.	Prediction algorithms, Industries, Business, Random forests
2.	Y. Parikh and E. Abdelfattah, "Clustering Algorithms and RFM Analysis Performed on Retail Transactions," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020, pp. 0506-0511, doi: 10.1109/UEMCON51285.2020.9298123.	Data mining, pattern clustering, purchasing, clustering algorithms,High frequency, retail data processing
3.	D. Bin, S. Peiji and Z. Dan, "Data Mining for Needy Students Identify Based on Improved RFM Model: A Case Study of University," 2008 International Conference on Information Management, Innovation Management and Industrial Engineering, 2008, pp. 244-247, doi: 10.1109/ICIII.2008.128.	Data mining, frequency, information analysis, databases, information management

4.	A. Miklosik, M. Kuchta, N. Evans and S. Zak, "Towards the Adoption of Machine Learning-Based Analytical Tools in Digital Marketing," in IEEE Access, vol. 7, pp. 85705-85718, 2019, doi: 10.1109/ACCESS.2019.2924425.	Decision making, internet, strategic planning, tools
5.	S. Bouktif, A. Fiaz and M. Awad, "Augmented Textual Features-Based Stock Market Prediction," in IEEE Access, vol. 8, pp. 40269-40282, 2020, doi: 10.1109/ACCESS.2020.2976725.	Feature extraction, Sentiment analysis, predicting models, machine learning
6.	M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," in IEEE Access, vol. 7, pp. 106111-106123, 2019, doi: 10.1109/ACCESS.2019.2930410.	Optimisation,Big data, data visualization, data mining.
7.	Yi Wang, Qixin Chen, Chongqing Kang, Mingming Zhang, Ke Wang and Yun Zhao, "Load profiling and its application to demand response: A review," in Tsinghua Science and Technology, vol. 20, no. 2, pp. 117-129, April 2015, doi: 10.1109/TST.2015.7085625.	Demand response, data mining, load management,feature extraction

8.	I. Khan, J. Z. Huang, M. A. Masud and Q. Jiang, "Segmentation of Factories on Electricity Consumption Behaviors Using Load Profile Data," in IEEE Access, vol. 4, pp. 8394-8406, 2016, doi: 10.1109/ACCESS.2016.2619898.	Data mining and clustering algorithms.

2.2 Existing System

To make developing deals easier, cultivate relationships with current clients, and don't let client relationships sour after the initial transaction. Learn how to increase revenue from existing clients:

A. Increasing revenue from existing clients
B. Using marketing automation to build long-term relationships with customers
C. Respond to changing client requirements
D. Develop new products that are targeted towards your current customers.
E. Rather than focusing just on new customer acquisition, upsell and strategically pitch as part of your standard advertising strategy.
F. Make a program for trustworthiness.
G. Create a team which is entirely focused on the clients.
H. The overall worth of a client

1. Investigate the market.

Statistical research and analysis provide doors to new possibilities inside your current consumer base. You can focus on who your clients are and what they need by researching the market you service.You can learn more about your customers by using a variety of sites, such as:

A. Directing a market investigation to gather information.
B. Research, center gatherings, and meetings are all used to gather feedback.
C. Throughout the friendlier stages, paying attention to what clients say.
D. Go through the trade magazines and the correspondence.

Also look at data from the Bureau of Labor Statistics and the Economic and Statistics Administration. Learn about important components for example attribution across numerous channels using the Google Analytics data. AOV section and regional features have been distributed. On your website, client journeys are also divided by segment and area. You may use it to tell the tale of your image through web-based media, email, live chat, and any other channels where your clients may tune in. Get to know more about the client's venture planning here. Keywords which attract most of the traffic, similar to what one can look for in the circumstance.

Set objectives that are comparable to the return on investment (or, more accurately, the return on advertising expenditure) of different initiatives on your website to calculate the return on investment (ROI) from each channel and marketing message, the appearance of your welcome pages, and many website components' devotion Concentrating on your market will allow you to examine your current clients' thought patterns, common buying paths (ventures), tangles in the client venture, and so on and, in any case, identifying overlooked requirements that reflect opportunities for you to differentiate yourself from the competition.

2. Trade models that are compatible.

Customers have many choices, but still if they choose you then it's just wonderful even after that tough competition all over. However, after you've completed the initial transaction, stay in touch to form mutually beneficial ties. Rather than merely chasing another sale, it's a good idea to follow up with existing clients by delivering them something helpful.

Current clients should be approached with low-pressure conversations. You've already persuaded the customer to buy from your company. At this point, you must unquestionably continue the conversation.

Advertising automation, such as Salesforce or Marketo, and the others mentioned above, monitors client relationships over time by sending tailored messages based on previous interactions with individual customers/clients. As a result, if you release a new item or update an existing one, such as one that a client previously purchased, the client will be notified. This information is provided through a custom message.

At the retail location, you gathered client contact information. Most rules allow you to save the information for future correspondence. Don't waste that opportunity by sending out follow-up deals notifications until your clients opt out of future communication. On friendly venues like Instagram, use instruments like email, calls, and hollers.

2.3 Proposed system

The major recommended system that we employed here is more focused on new tools and technologies that have been released in recent years. They include technology such as data science, data visualization, data analyzation, machine learning, deep learning,and others.

To fuel the growth of any firm, we followed the data science method of first locating the dataset, evaluating it, and then utilizing the model we generated after training and testing.

Instead of relying solely on humans, the proposed system automates the majority of the analysis, which is accomplished using Jupyter Notebook, Python libraries that we imported such as Kera, tensorflow, numpy, panda, scikit, and xgboost, and using these libraries to automate and build a model to visualize how the change is occurring in the given dataset.

2.4 Feasibility Study

Any necessary stage of the product development process has been obtained. Allows engineers to obtain a tested and functional object. Alludes to item investigation in terms of item results, application execution, and specialized assistance required to use it. A prospective examination should be done based on a variety of factors and circumstances.

2.4.1 Economic Feasibility

The difference between the benefits or results we gain from an object and the overall cost we pay to enhance it is referred to as monetary recovery. In the existing framework, the development of additional items enhances framework accuracy and speeds up application and announcement handling.

2.4.2 Feasibility Probability

Accessibility refers to the display of an object in order for it to function. While some things may operate admirably throughout production and use, they may break down in reality. It entails looking into the necessary characters as well as their specific information. In the current architecture, the contained information, updated statistics, and reports for ages are precise and quick.

2.4.3 Technical Considerations

The term "specialized execution" refers to whether or not the currently available product can fully support the present framework. It considers the benefits and drawbacks of employing explicit advancement programming, as well as its applicability. It also calculates how much more time clients will need to make the programme work. The current framework's user interface is simple to use and requires little information or preparation. It only takes a few mouse clicks to complete exercises and generate reports. Customers require quick access to sites with an unquestionable level of security, thus the solution used to update is ideal for current applications. This is accomplished by combining a web server and an information server at the same physical location.

CHAPTER 3: SYSTEM DEVELOPMENT

3.1 Design and development

In any case, most importantly, why do we do division?
Because you can't provide each client with a similar matter, channel, or priority. They will discover different services that provide exactly what they want and they feel more connected or go to brand for them.

Method I used in this project are the Elbow method:

Elbow Method : This Method is among the most often used methods for calculating the ideal value of k. The number of clusters that is ideal into which the sample points can be separated is crucial in any unsupervised method. You can utilize K-means clustering in the customer segmentation module for getting the value of recency.We should, however,Inform the K-means technique of the number of times groups you require. We used the Elbow Method to check this out. For maximal inertia, the Elbow Method simply informs us how many clusters there are.

Following are some key definitions and explanations:

Naive Bayes is a classification technique that can categorize binary and multiclass data.. It is a supervised classification technique that uses conditional probability to assign class labels to instances/records in order to categorize future objects. It also plays an important role in this project just like XGBoost.

Hyperparameter tuning: A mathematical model containing a number of parameters that must be learned from data is referred to as a Machine.
Hyperparameters, on the other hand, are a special form of parameter that cannot be learned using a regular training approach.They are normally fixed prior to the start of the training

procedure. These values describe important features of the model, such as its complexity and learning rate.

EDA (Data Exploration Analysis): EDA is a data assessment strategy that employs a variety of (mostly diagrammatical) methods to optimize comprehension of a data set. This apart from fitting the infrastructure to available information, we can fit the same parameters of the model. by building the classifier with existing data Recognize underlying structure, extract significant factors, Identify outliers and anomalies, investigate underlying assumptions, build parsimonious models, and determine the best factor settings.

3.2 Model Development

Dataset used

Dataset which I used in this project is Online Retail from kaggle. Using this dataset I analyzed and visualized and tried to come up with better results and models in lifetime value of customers module. This dataset includes many important features in itself which includes like

- Id of the customers
- Prices of each unit
- Date of bills

And some other features are also there in that dataset. Using all this one can calculate the revenue, which is really important in this project as it helps to understand other factors more clearly.

I'll create a simple machine learning model that anticipates the value over time of our clients. This is how we're going to get there:

- Define an acceptable time range for calculating clients LTV.
- Recognize and create the features that will be used to make future predictions.
- Calculate the LTV of the ML model you're training.
- Create a ML model then run it.
- Check to see if that model is applicable.

The other dataset I used for marketing analysis includes columns that tell the customers personal information, columns showing the personal attitude or way of acting of consumers and actions or responses of consumers based on the different columns.The columns in that dataset involves ID, Year Birth, Education, Marital status, Income, Kidhome, Tennhome, Dt Customer, Recency, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumDealsPurchase, NumWebPurchase, NumCatalogPurchase, NumStorePurchase, NumWebVisitsMonth, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, AcceptedCmp2, Response, Complaint, Country. Many of these features are pretty important in order to calculate the important factors and for the visualization and the results which I got with the great efficiency itself. So it feels like the best dataset for the marketing analysis part because it covers almost all the relevant variables in it. So using all these features/columns I visualized and analyzed the results and tried to come up with the best possible conclusions and results.

So, basically these two dataset I've used in this project and they both are from Kaggle only and I feel that they are highly suitable because of all the reasons mentioned above with the explanations.

3.3 Mathematical

Mathematical formula which is used in the project is:

- Total Gross Revenue - Total Cost = Lifetime Value

We can calculate Lifetime Value for each customer using this calculation equation.

- Revenue = Active Customer Count * Order Count * Average Revenue per Order

Used this formula to calculate the revenue as we consider revenue as our north star metric for the dataset on internet sales.
These are some of the formulae used in this project directly.

3.4 Tools and Technologies used

- Anaconda : Jupyter Notebook
- Python Programming language
- Libraries like numpy, pandas, scikit, matlab, pyplot etc.
- Machine learning models/algorithms
- Data science process

A quick rundown of the libraries and packages utilized here:

Numpy is a Python library that allows you to handle arrays. Numpy might even try to deal with algebraic expressions and direct advanced math. NumPy is the abbreviation for Numerical Python.

What is NumPy and why should you use it? So, in Python, we have lists that function similarly to arrays, but they are slow to process. NumPy is made to make array objects 50 times faster than ordinary Python lists.
ndarray in NumPy's array object, and it comes with a number of auxiliary functions to make collaboration easier. Arrays are commonly used in data research when speed and resources are important.

Pandas is a well-known and well-liked data science application that uses the Python computer language to manipulate and analyze data. In the real world, samples/data is inherently chaotic. In case of cleaning, manipulating, altering, and analyzing data, Pandas is a game changer. In essence, this aids in the complete cleanup of the waste.

Matplotlib:Matplotlib is a GUI for Python and a diagrammatical charting programme that is a statistically upgraded version of NumPy. It serves as a free software alternative to MATLAB.

Seaborn is a visualization kit for the Scripting language that is also based on Matlab. It includes a high-level interface for building aesthetically appealing and educational statistics visualizations.

Plotly: Plotly allows users to import, copy and paste, or stream data in order to examine and visualize it. Python scripts can be saved, shared, and collaborated on with Plotly.

Datetime: The datetime module is responsible for manipulating dates and times.

Sklearn (Computational tool) is the most functional and dependable open source pattern recognition library. It provides a variety of machine learning features using a Python consistency interface. classification, regression, clustering, and dimensionality reduction, as well as statistical modeling.

CHAPTER 4: PERFORMANCE ANALYSIS

4.1 Discourse: Finding Results

Here's I’ll be explaining the LTV as:

- ➔ Customer lifetime value prediction: So here I can compute the LTV of the consumers. That's kind of one of the essential factors or the metric. To create income and to profit-making, a company invests a lot on their consumers through many ways which includes digital and offline advertising and advertisements, discounts or interesting deals and some other ways also. Automatically all this kind of action will lead to improving the LTV of the customers in a drastically manner if done rightly and in a smart way but even after all the attempts by the company, still some clients will be there which will lead to the disappointment of those efforts. So, for that to avoid or to waste the effort on the wrong client, they should come up with the solutions to it like they can segregate their consumers appropriately, and analyze trends and responses individually and take appropriate action for better results.

So here I am, constructing the ML forecasting model that effectively checks the lifelong value of customers. First, let's figure out how to accomplish it properly and sequentially:

- Set an acceptable time frame for computing Consumers LTV.
- Find and develop the traits that will be used to make predictions for the future.
- Obtain an ML model's LTV.
- Build a ML predictive model, then run that.
- Check to see if the model is adequate.

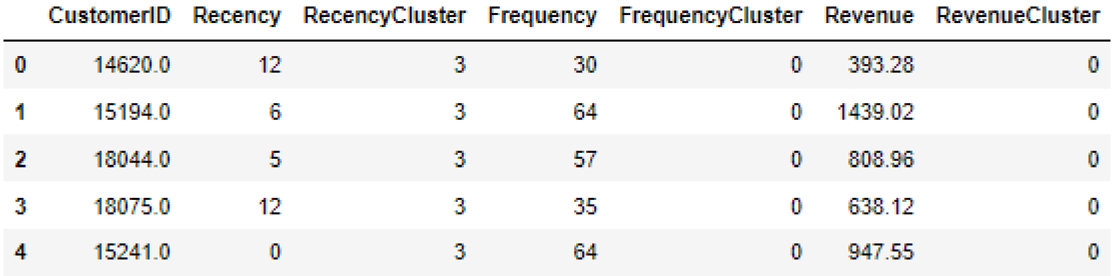
For measuring LTV, we must select a time frame. Three, six, twelve, or twenty-four months are possible options. It all varies based on the type of industry/company, the marketing strategy used, as well as the vision, vision, and other elements. For some businesses, a year

can be a lot longer, while for others, it can be a quite brief time. This dataset would be 6 months long therefore in this situation. We may measure Overall Value for each consumer all through that period span using technique below:
LTV = Total Earnings Revenue - Overall Cost

It is now possible to determine the previous lifespan amount while using formulas. It may be too difficult to intervene if we discover that some clients have a rather significant negative overall worth. Machine learning should be used to forecast however at time.

RFM scores for each customer ID provide nice option group contenders. We must segment the information to completely implement it. We'll take three months' worth of data, estimate RFM, and use that to forecast the next 6 months. As a result, make two dataframes and start by adding RFM values to these.

The following features are now available after finishing the RFM scoring:



	CustomerID	Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue	RevenueCluster
0	14620.0	12	3	30	0	393.28	0
1	15194.0	6	3	64	0	1439.02	0
2	18044.0	5	3	57	0	808.96	0
3	18075.0	12	3	35	0	638.12	0
4	15241.0	0	3	64	0	947.55	0

Table: RFM Scoring

This is the dataset, therefore, as i have the dataset ready, now i'll calculate the ltv of the consumers. As above, I decided the 6 months as the time frame so here I'll be Generating the six month lifespan value, which would be important throughout the ML model's training.

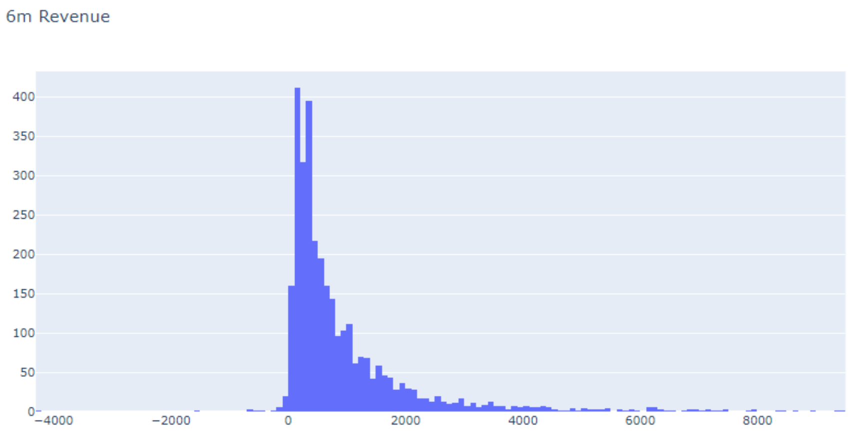


Fig: 6m Revenue

We can deduce from this histogram that there are consumers with a negative LT value. After removing the outliers, I combined the 3- and 6-month characterization studies seeing the connection among LTV and the features we're looking at.

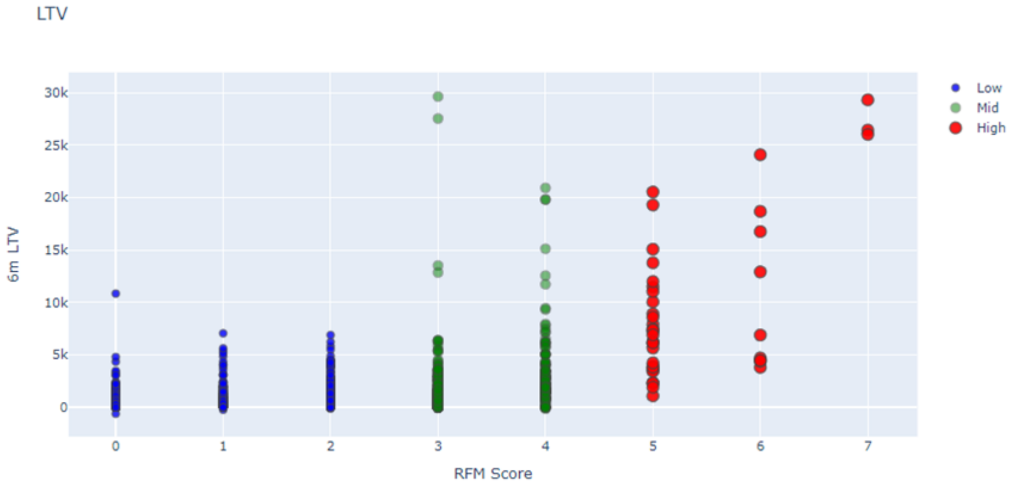


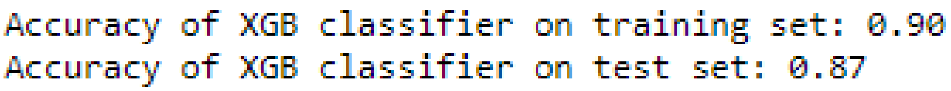
Fig: LTV vs overall RFM score

From this result we could see that the RFM scores and LTV have a positive relationship.

Keeping the commercial aspect in mind, you must differentiate the consumers in order to succeed and here in this case i am differentiating them on LTV scores basis. So, here i'll be using the clustering and simply cluster them or differentiate/segment them into clusters as:

- Low LTV
- Mid LTV
- High LTV

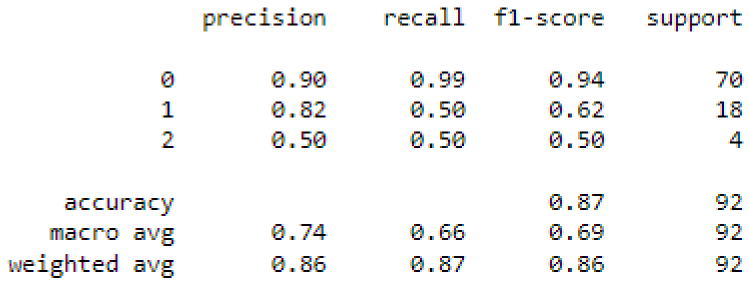
I utilized XGBoost, a sophisticated ML toolbox, and classify. It grew into a multi-category model even though we had three groupings (clusters).



```
Accuracy of XGB classifier on training set: 0.90
Accuracy of XGB classifier on test set: 0.87
```

Table: Accuracy

Finally, we can observe how I got a 90% correctness here on the training dataset and an 87 percent accuracy on the test set.



```
              precision    recall  f1-score   support

           0       0.90      0.99      0.94        70
           1       0.82      0.50      0.62        18
           2       0.50      0.50      0.50         4

    accuracy                           0.87        92
   macro avg       0.74      0.66      0.69        92
weighted avg       0.86      0.87      0.86        92
```

Table: Classification report

For a value of 0, precision and recall are acceptable. If the model indicates this customer belongs to cluster 0 (low LTV), 90 out of 100 times it will be right (precision). Furthermore, the model successfully identifies 99 percent of actual cluster 0 customers (recall).

We now have a machine learning model that forecasts future LTV categories for our consumers. As a result, we may easily change our activity. For example, we don't want to lose customers with a high LTV.

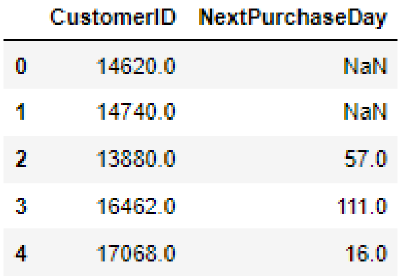
- ➔ Predicting the clients' next purchasing day: Serve your clients with respect, taking into account the lifetime worth, corporate data, and taking action now before it negatively occurs, such as desertion (customers start churning).
- There are numerous methods that a corporation can take to avoid any of these issues by simply improving its current condition, including:
- Consumers who purchase through the firm in all situations will not receive any form of discount.
- Use brand awareness to push the buyer if they do not make a purchase within the expected time range.

I've used the same dataset that I've used in the LTV modele.
Under this module, these are the steps that I followed:

1. Data manipulation (building previous / next records and determining overall gap among buying days)
2. Characteristics Engineering
3. ML Model Selection
4. Hyper - parameters Adjustment System for Multi-Classification

Data wrangling: Import the csv file first, then transform the data field. To construct a model, we should separate our dataset in two portions.
Based on six months of observations, forecast the client's initial buying day over the next three months. Will also forecast even though no product is bought. Suppose that 9th of September as the cut-off day, and divide the dataset as follows:
Tx 6m represents six-month achievement, and tx next represents those dates in between the last buying day in tx 6m and the first purchase date in tx next.

I'll be generating a dataset called tx user, which will supply the forecasting models with a consumer set of features:

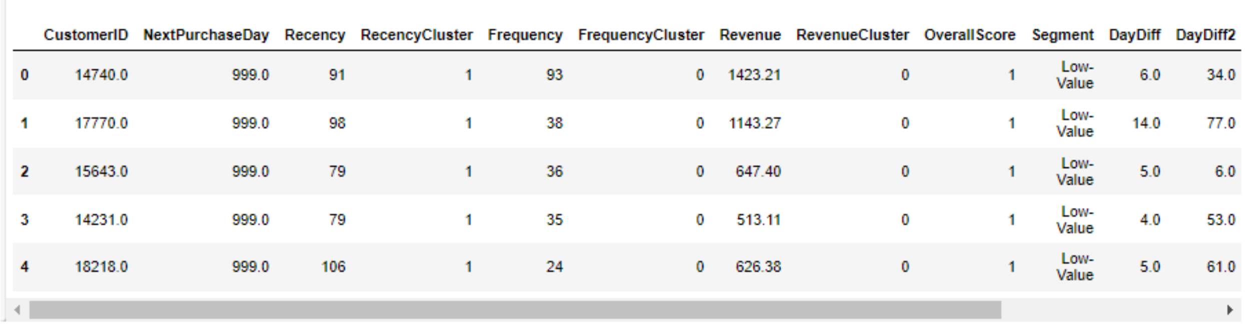


	CustomerID	NextPurchaseDay
0	14620.0	NaN
1	14740.0	NaN
2	13880.0	57.0
3	16462.0	111.0
4	17068.0	16.0

Table: tx_user

The data frame is made up of the consumers' unique identifiers as well as their labels. Now it's time to start adding characteristics to the ML model.

- Engineering characteristics: The following are the features that were selected to construct that model:
- RFM ratings and groupings
- How long was that before or after your previous three buyers?
- The average ± STD of the days among buys diverge



	CustomerID	NextPurchaseDay	Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue	RevenueCluster	OverallScore	Segment	DayDiff	DayDiff2
0	14740.0	999.0	91	1	93	0	1423.21	0	1	Low-Value	6.0	34.0
1	17770.0	999.0	98	1	38	0	1143.27	0	1	Low-Value	14.0	77.0
2	15643.0	999.0	79	1	36	0	647.40	0	1	Low-Value	5.0	6.0
3	14231.0	999.0	79	1	35	0	513.11	0	1	Low-Value	4.0	53.0
4	18218.0	999.0	106	1	24	0	626.38	0	1	Low-Value	5.0	61.0

Table: Grouping the label

The feature set is complete and ready to be used in the creation of a classification model. But, with so many models to choose from, how can we know which one to use?

How to Choose a Machine Learning Model: Firstly take two steps before deciding on the model. First, we must determine which classes our label contains. Percentiles, in general, provide the proper answer. Let's use this as an example. To see them in NextPurchaseDay, use the describe() method:

The question of determining the boundaries applies to both statistics and commercial demands. It should make sense in relation to the first and be simple to implement and express. Having three classes based on these two:

- Customers who will buy in the next 0–20 days — Class name: 2
- Customers who will buy in the next 21–49 days — Class name: 1
- Customers who will buy in the next 50 days — Class name: 0

The final step is to examine the relationship between our characteristics and the label. One of the simplest methods to demonstrate this is with a correlation matrix:

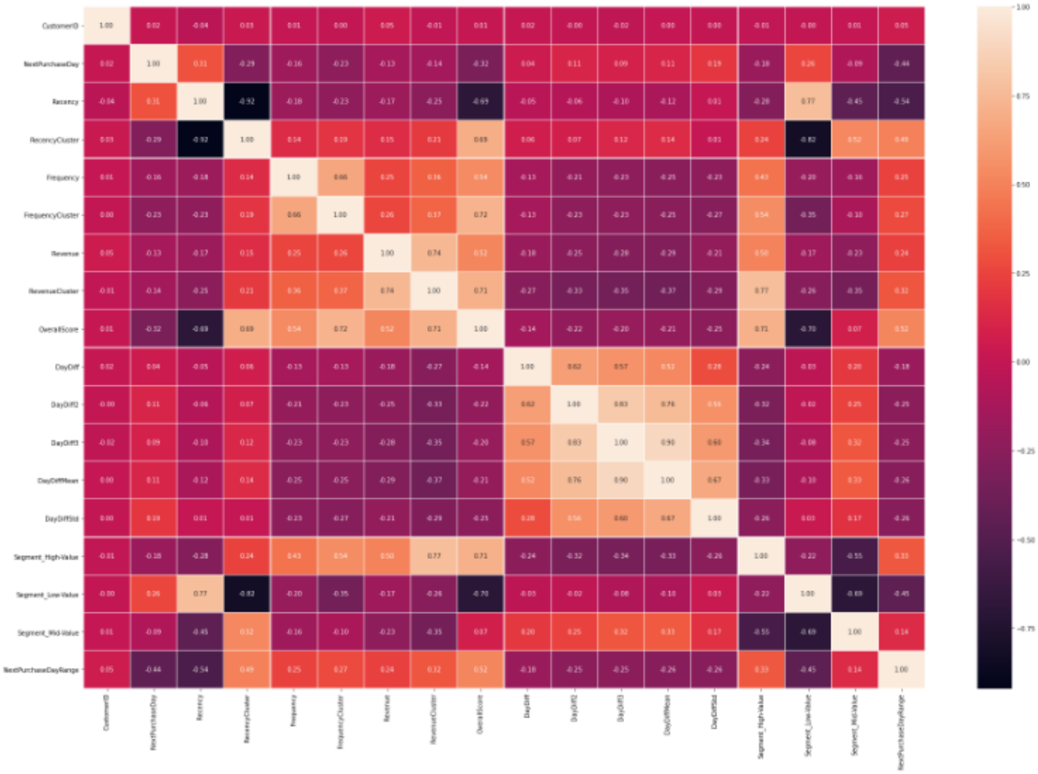
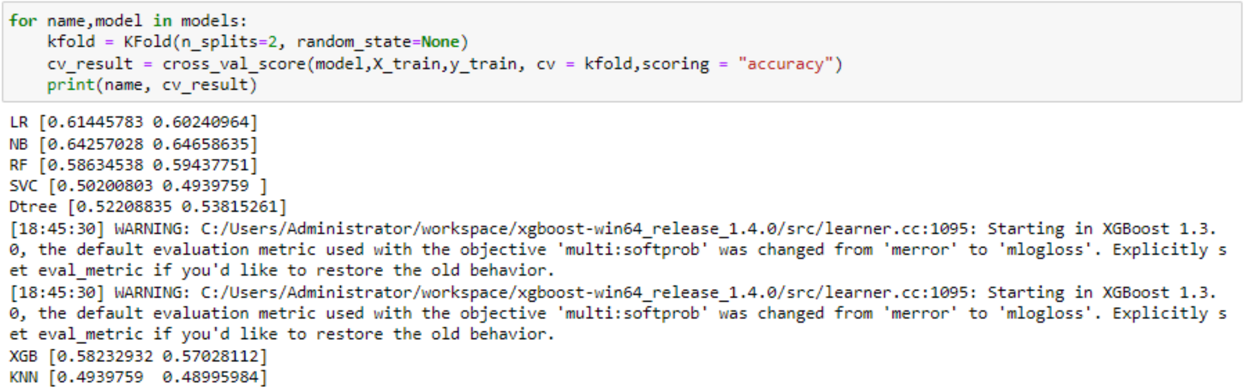


Fig: Correlation

Recency seems to have the biggest negative relation, while Overall Score seems to have the strongest positive connection (0.45). (-0.45). (-0.54).

For this task, we want to use the model that delivers the maximum correctness. Let's examine the correctness of several models by separating the train and test phases:



```
for name,model in models:
    kfold = KFold(n_splits=2, random_state=None)
    cv_result = cross_val_score(model,X_train,y_train, cv = kfold,scoring = "accuracy")
    print(name, cv_result)
```

```
LR [0.61445783 0.60240964]
NB [0.64257028 0.64658635]
RF [0.58634538 0.59437751]
SVC [0.50200803 0.4939759 ]
Dtree [0.52208835 0.53815261]
[18:45:30] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.
0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly s
et eval_metric if you'd like to restore the old behavior.
[18:45:30] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.
0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly s
et eval_metric if you'd like to restore the old behavior.
XGB [0.58232932 0.57028112]
KNN [0.4939759  0.48995984]
```

Table: Accuracy of models

In the end, I achieved a Naive Bayes model with a 64 percent correctness in predicting the consumer's future buying day.

➔ <u>Marketing Analysis:</u>

The stages in the market research method are as follows:

1. First and foremost, I must examine and sanitize the data.
2. Analysis of data and investigation
3. Analytical Analysis of data
4. Visualizations and Interpretation on a Large Scale
5. Developing Data-Driven Solutions

Step 1: Information Extraction and Evaluation

Consider the feature first. The first 11 columns provide the following information about the customer:

- ID: The customer's individual identifier.
- Year of Birth: A consumer's newborn date.
- Customer Education: The customer's degree of knowledge.
- Customer marital status: The customer is married or not.
- Income: The customer's annual revenue.
- Kidhome: The total number of kids in the customer's household.
- Tennhome: The amount of adolescents who live with the consumer.
- Dt Customer: Registration time for the consumer.
- Amount of days after the last buy by a consume

The following 11 columns detail the behavior of consumers:

- MntWines: An overall sum spent buying alcohol over the previous twenty-four months.
- MntFruits: Fruit purchased with in previous two twenty-four months.
- MntMeatProducts: The money spent for meat throughout the previous two years.
- MntFishProducts: The overall sum spent buying fish throughout the previous two years.
- MntSweetProducts: Sweets bought throughout the preceding two years
- MntGoldProds: The expenditure incurred for gold throughout the previous two years.
- numDealsPurchase: Total amount of discounted purchases
- numWebPurchase: Amount of items purchased by company's website
- NumCatalogPurchase: The overall number of catalogs buys.
- NumStorePurchase: The overall amount of store items bought.
- NumWebVisitsMonth: The amount of visitors to a company's website throughout the previous month.

Customers' answers to earlier campaigns are collected in these unused columns:

- AcceptedCmp: If the offer was accepted in the first campaign, AcceptedCmp1 is 1, otherwise it is 0
- AcceptedCmp2: If the offer was accepted in the first campaign, AcceptedCmp2 is 1, otherwise it is 0
- AcceptedCmp3: 1 whereas if third marketing deal was done, 0 otherwise.
- AcceptedCmp4: 1 if the offer was accepted during the fourth campaign, 0 otherwise the.
- If the consumer accepted the offer in the fifth campaign, Cmp5 is 1, otherwise it is 0.
- If the buyer accepted the prior campaign's proposal, 1; else, 0.
- If a client has filed a complaint during the last two years, 1; otherwise, 0.
- Country: Customer's country of residence.

This dataset contains 28 columns, 2240 rows, and zero duplicate records.
The EDA is the following stage after fully accessing and cleansing the data.

Step 2: Data Analysis for Research

In this stage 2, I'll focus on the following three questions:

1. Is there anything that stands out or that is unusual?
All of the numerical features were shown using boxplots, and the data's five numbers will be displayed: the lowest non-outlier number, Q1(25th percentile), Q2(50th percentile), Q3(75th percentile), and the highest non-outlier number.
So, below are the boxplots using which I can check all the outliers correctly.

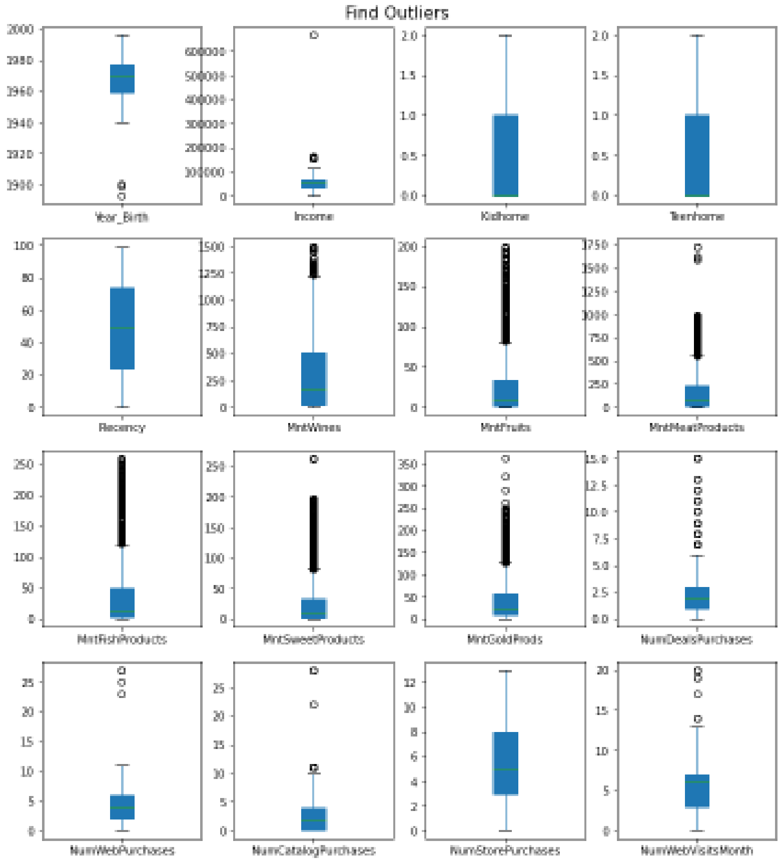
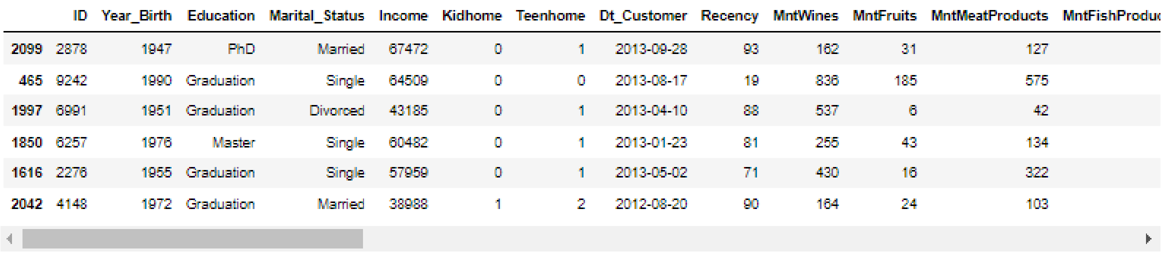


Fig: Box Plots (Find Outliers)

Outliers appear in multiple columns, but the majority of them appear to be natural anomalies. These anomalies in Year of Birth, on the other hand, appear to be entry errors, as no one alive today could have been born before 1900.

2. Is there any way to make relevant variables using the data you have?



	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProduc
2099	2878	1947	PhD	Married	67472	0	1	2013-09-28	93	162	31	127	
465	9242	1990	Graduation	Single	64509	0	0	2013-08-17	19	836	185	575	
1997	6991	1951	Graduation	Divorced	43185	0	1	2013-04-10	88	537	6	42	
1850	6257	1976	Master	Single	60482	0	1	2013-01-23	81	255	43	134	
1616	2276	1955	Graduation	Single	57959	0	1	2013-05-02	71	430	16	322	
2042	4148	1972	Graduation	Married	38988	1	2	2012-08-20	90	164	24	103	

Table: Relevant Variables

As a result, the most important factors can be carried out from here. So those crucial variables are Join month, Join weekday, Minorhome, Total Mnt, Total num purchase, Total accept, and AOV.

3. Is anything unusual about info?

These relationships among each parameter can be seen using a heatmap. When it grows bluer, they become more positively correlated, and when it gets redder, they become more negatively correlated.

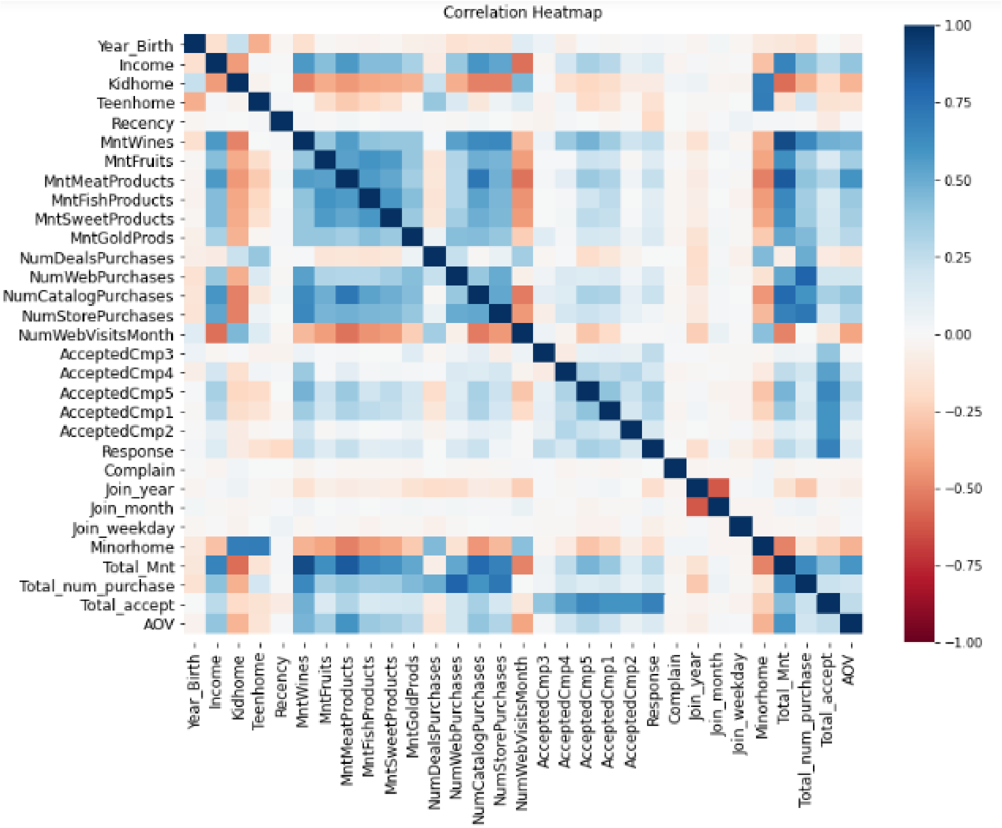


Fig: Correlation Heatmap

Findings:

1. High-Income Individuals People who spend and buy more visit the company's website less frequently than those who only make a few discounted purchases.
2. People who have children at home spend less and buy less, with a large percentage of their purchases being made at a discount.
3. People who bought a lot of things from the catalog tended to buy more alcohol and meat stuff and didn't go to the company's website.

I anticipated that the extra complaints a client has, the less likely they are to spend in our store; however, the amount of complaints throughout the past couple of years has no correlation with the overall spent in the previous two years.

After going deeper into the data, I realized that it's because only 20 consumers out of a total of 2200 have complained in the last two years. Because of the unequal ratio, they don't match. The company's customer service workforce has performed admirably in the previous two years.

Step 3: Conduct the necessary calculations.

Within that step, I'll mostly look into the three possible questionnaire:

1. Which factors influence how much money people spend in stores?

We may use the random forest to forecast shop purchases and then use the model's feature importance score to rank the components.

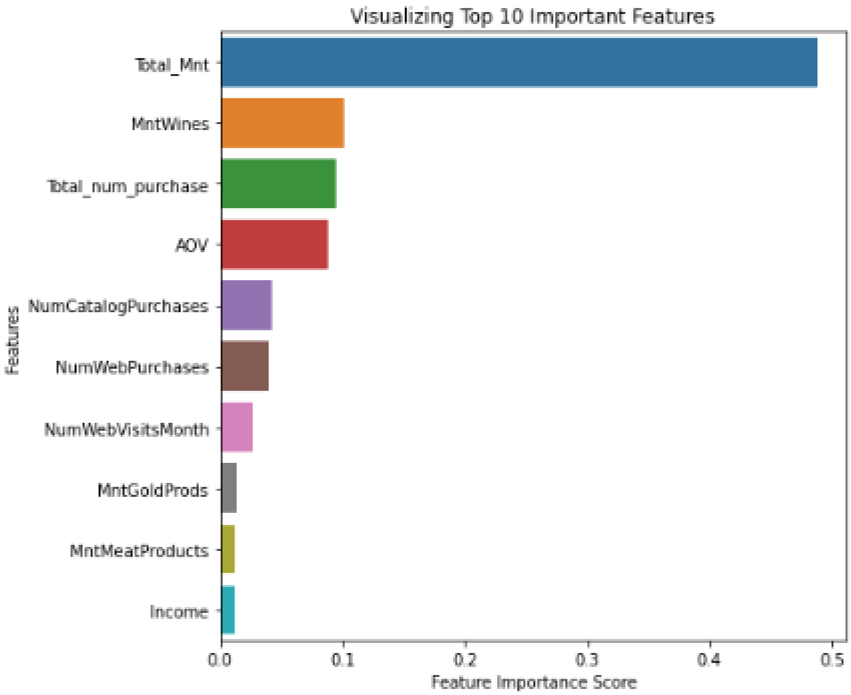


Fig: Visualizing top 10 important features

As can be seen, the top seven factors are:

1. Average order volume
2. Total spending over the previous two years
3. Purchases done within the last two years
4. Wine spending during the past two years
5. Total amount of catalog orders
6. The number of visits to the company's website in the previous month
7. In the previous two years, the total number of internet purchases

However, we can't determine if each factor has a positive or negative impact on the volume of retail transactions. SHAP can be used to describe it.

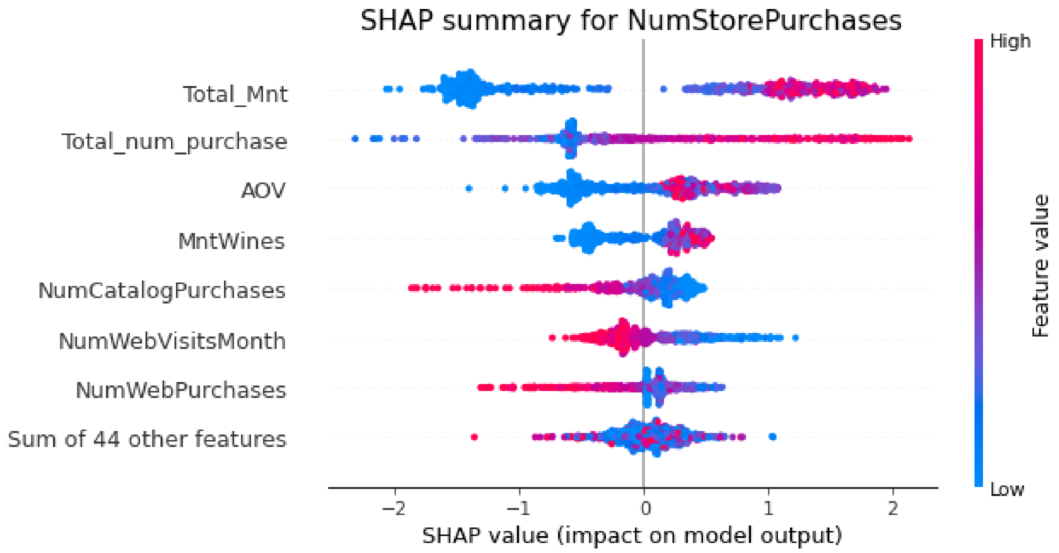


Fig: SHAP summary for NumStorePurchases

People who mostly shop in stores buy more wines, have a higher average order volume, and utilize the internet and catalogs less.

Finding:

A. The number of shop purchases increases as the overall expenditure made (Total Mnt), final amount value (Total num buy), AOV, & quantity of wines purchased increase (MntWines).
B. As the amount of web visits (NumWebVisitsMonth), portfolio sales (NumCatalogPurchases), and website purchases increase, the number of retail purchases decreases (NumWebPurchases).

2. The belief that people who buy gold are more conservative. As a result, customers who spent more money on gold in the previous two years would make more purchases in-store. To prove or disprove this claim, use a stat test.

A corr test can be often used to evaluate to check if MntGoldProds and NumStorePurchases are directly associated. Let's start with the two-variable frequency distribution.

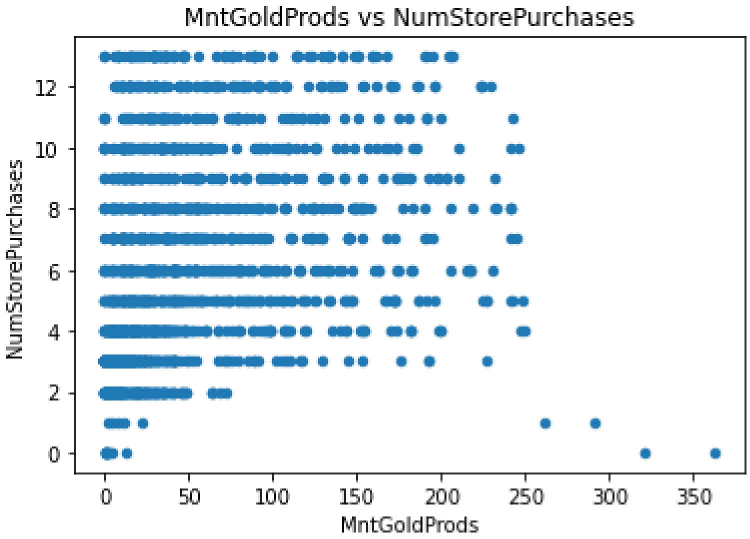
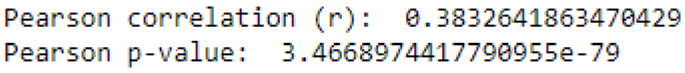


Fig: MntGoldProds vs NumStorePurchases

So, seeing the diagram we could conclude that if MntGoldProds increase, the association in MntGoldProds and NumStorePurchases is getting unclear.

So, it's time to take a correlation test here.



```
Pearson correlation (r):  0.3832641863470429
Pearson p-value:  3.466897441779 0955e-79
```

Table: Correlation Test

I received a Pearson correlation of 0.38 and a pearson p-value of nearly zero from the above, showing that they are statistically significant and have a positive relationship. (If the p-value is more than 0.05, the null hypothesis that they are unconnected cannot be rejected.)

3. Fish oil contains omega 3 fatty acids, which are good for the brain. Is there a link between the amount of money spent on fish and the number of married Ph.D. candidates?

To statistically validate these, I divide the data into two groups. One is the married Ph.D. group, and the others are the rest. Then, using a boxplot, we can compare the two groups to see if they vary. Finally, a t-test can be used to determine whether their means are equivalent.

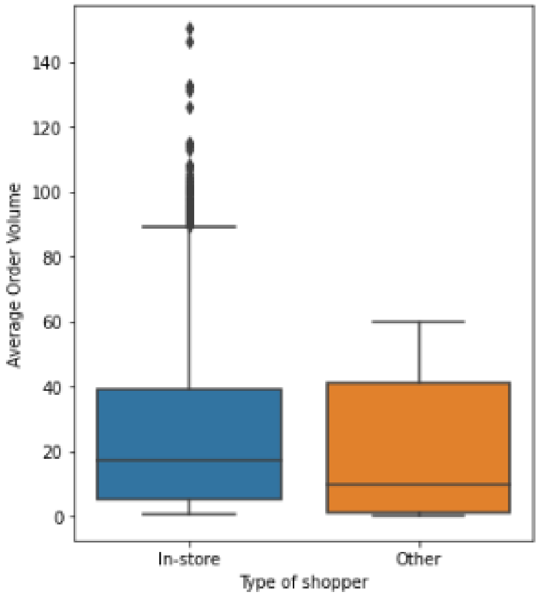
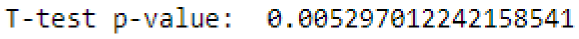


Fig: Married PHD vs the rest

As seen by the fact that the 50th percent stands more than what the married Ph.D. group, the rest of the consumers spend more on fish goods. Let's look at the t-test right now.



```
T-test p-value:  0.005297012242158541
```

Table: T-test

This h0 is rejected because the p-value was little below 0.05, implying their means are different, though the Married Ph.D. 's mean is lower than the others, as seen in that figure.

Step 4: Visualization of the data and Additional Research

The questionnaires I'll cover here as written below:

1. Which promotional strategy works the best out of all other ones?

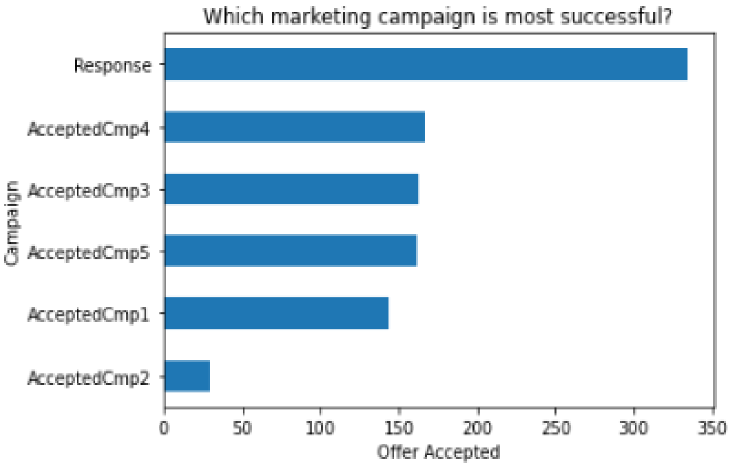


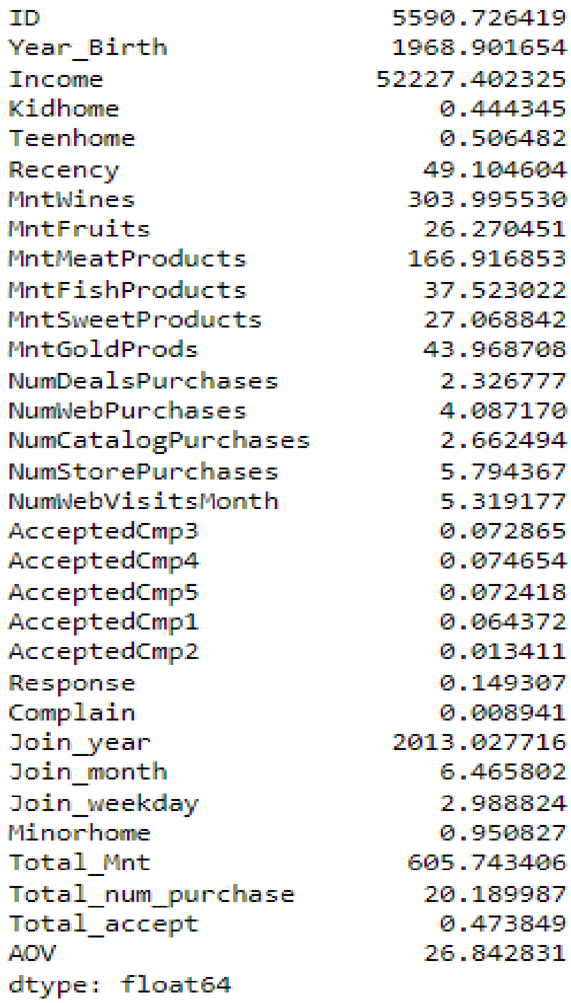
Fig: Most successful campaign

The far more recent and effective business initiative is referred to as reaction. With the exception of initiative 2, it performed almost twice as well as that of the previous initiative.

2. Who is the typical shopper for all of this company? Is there one of the most effective items?

Using .mean(), I observed that a typical consumer looks like this:

- Gets a yearly salary of 52227.40 dollars
- Has an AOV of 26.84 dollars and was purchased 49 days ago
- 605 dollars has been spent
- Has made 20 purchases
- In mid-June, I became a client.
- On Thursday, I became a client.
- The maximum amount spent on wines ($303) and then meat goods (166 dollars)
- Fruit and sweet items were the least expensive (26 dollars) (27 dollars)



```
ID                     5590.726419
Year_Birth             1968.901654
Income                52227.402325
Kidhome                   0.444345
Teenhome                  0.506482
Recency                  49.104604
MntWines                303.995530
MntFruits                26.270451
MntMeatProducts         166.916853
MntFishProducts          37.523022
MntSweetProducts         27.068842
MntGoldProds             43.968708
NumDealsPurchases         2.326777
NumWebPurchases           4.087170
NumCatalogPurchases       2.662494
NumStorePurchases         5.794367
NumWebVisitsMonth         5.319177
AcceptedCmp3              0.072865
AcceptedCmp4              0.074654
AcceptedCmp5              0.072418
AcceptedCmp1              0.064372
AcceptedCmp2              0.013411
Response                  0.149307
Complain                  0.008941
Join_year              2013.027716
Join_month                6.465802
Join_weekday              2.988824
Minorhome                 0.950827
Total_Mnt               605.743406
Total_num_purchase       20.189987
Total_accept              0.473849
AOV                      26.842831
dtype: float64
```

Table: Average customer

3. Examine the dissimilarities in client attributes and buying behaviors between the most effective campaign and the rest.

We can investigate the variations in individual customers and shopping habits between more promotional events, the last, and some others, promotions 1–5.

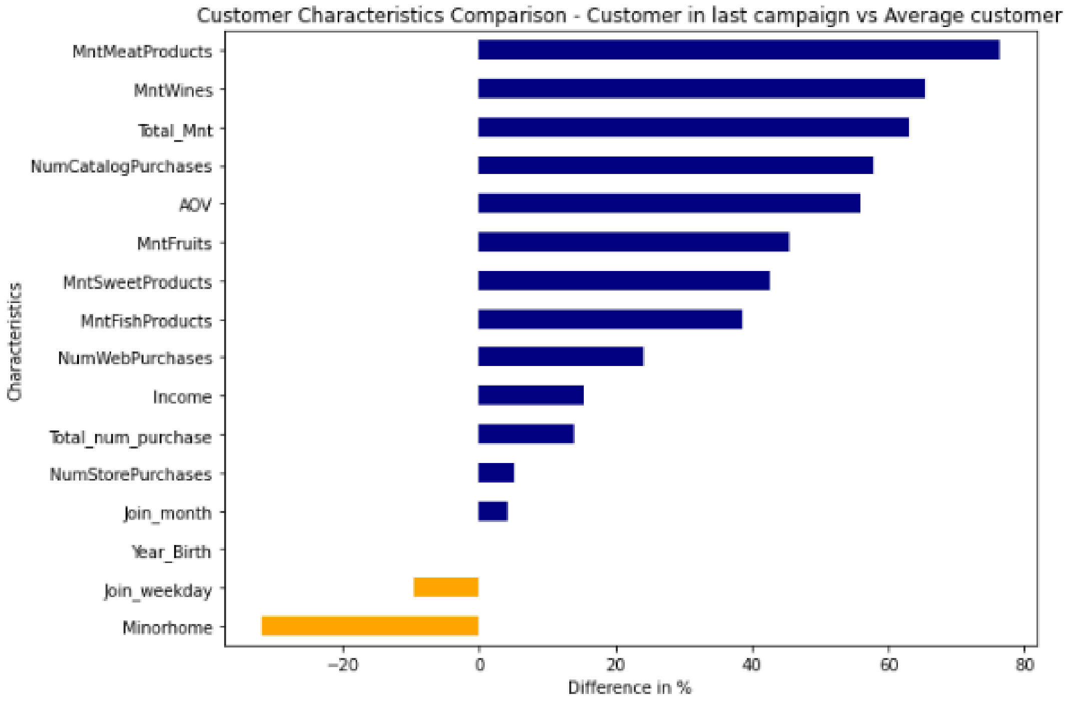


Fig: Customer in last campaign vs average customer

Step 5: Developing Data-Driven approaches

I have to research calculations in order to comprehend the issue and develop information alternatives. I combine all of the analytics findings and use them to develop real plans for developing data-driven alternatives.

The following is a summary of the findings:

1. The previous initiative was roughly twice as successful as the preceding ones.
2. Consumers tend to shop in person because they're more likely to spend so much per item. Consumers may have made more unplanned purchases after seeing similar shops.
3. When consumers buy in physical places, they are more likely to spend more money on each item.
4. The normal Thursday consumer.

Implementable data-driven approaches for acquirement:

1. Use the same promotional methods like the last initiative to sell meats and wines.
2. Spend more on advertising in Spain but spend less on advertisements in India.
3. Hold a product sale date on Thursday or a product discounted month in June to acquire new ones.

On raising profits:

1. Conduct promotional strategies to turn clients who typically buy from a site or catalog into in-store consumers, even though most in-store transactions get a good overall transaction amount.
2. Create a loyalty programme to retain as many high-income clients as feasible.

CHAPTER 5: CONCLUSION

5.1 Conclusions

To classify the data, I used XGBoost, a powerful machine learning toolkit. Since there are three groupings that are low ltv, mid ltv, high ltv, it evolved into a multi-category model (clusters) and I used XGBoost model here. And in that I got 87% accuracy on the test set and 90% on the train set. For a value of 0, precision and recall are acceptable. For example, if the model says this consumer belongs to cluster 0 (low LTV), 90 out of 100 will be correct (precision). In addition, the model correctly detects 99% of genuine cluster 0 consumers (recall).

So, now we have a machine learning model that basically predicts our customers' future LTV segments. We can readily adjust our activities in response to this. For example, we don't want to lose customers with a high LTV.
Basically using this model and analysis and visualization a company can target their customer accordingly and invest their money, energy, time and everything (resources) accordingly, which definitely improves the growth of the company/startup.

In prediction of the next purchase day of the customer in the final result the best model i got is naive bayes with 64% of accuracy. As in this part i tried many models and check their accuracy which came up as for XGBoost model it's 58%, logistic regression it is 61%, naive bayes it is 64%, random forest it's 62%, SVC it's 50%, decision tree is 51%, KNN is 49%. So, from here we can conclude that the naive bayes model is the best one among all. And even after this I did hyper tuning of the XGBoost model for the better result and then there I got 99% of the accuracy on the training set and 62% accuracy on the test set.

In marketing analysis the summary of findings here is that the last campaign was about twice as effective as the others before it. Consumers prefer buying in person because they are more likely to spend more money per item, they may have purchased more impulsively since they

saw similar things in stores, they are more inclined to spend more money on each item when they buy in physical stores, typical customers on thursday etc.
On acquisition, actionable data-driven solutions can be, continue to promote meat items and wines using the same marketing strategies as the last campaign, in Spain spend more money on marketing and less in India, to attract new customers, hold a brand discount day on Thursday or a brand discount in the month of June.
I learnt that the recent campaign is the best campaign till now. But still there's always room to improve therefore all one can do is in order to increase money generation in their company because most in-store transactions have a high average order volume, plan marketing stunts in order to convert customers who buy largely on a website or catalog to in-store purchasers. And create a loyalty programme to keep high-income clients as loyal as possible.

So, this is all about the results and the conclusion of all three parts (Predicting lifetime value of the customers, predicting next buying day of each consumer and the marketing analysis) of my project.

5.2 Future Work

The results show that the integration of multidimensional data with various classifications is possible. In this area, feature selection and dimensionality reduction approaches can be effective inference tools. Additional research is needed in this area to increase the performance of classification systems so that they can predict more variables.

I intend to parametrize the categorization systems in order to achieve high accuracy. I'm looking at a variety of datasets to see how Machine Learning algorithms may be utilized to increase retention rates, average revenue percentage, company influence on consumers, sales forecasting, and overall economy decision-making. I want to cut down on errors as much as possible while still keeping high accuracy. I hope to expand this project to include the ability to identify various sorts of datasets, such as sales revenue and north star metric, marketing campaigns-based datasets, as positive or negative long-term decisions for the consumer. On this front, developing an enhanced deep learning method employing KNN and other neural network libraries could be beneficial.

5.3 Applications

The xgboost model developed and libraries used from Machine learning and deep learning science to improve business making decisions for companies is relatively new.

It has a wide range of applications and can even help with the construction of more accurate machine learning models. When such powerful software or tools are used, they have the potential to transform the IT industry and business economic industries and transform them for the better by automating.

The economy is putting pressure on businesses and organizations to accelerate development and increase revenue.

Quickly developing programming as-a-administration (SaaS) organizations need to support twofold digit development rates to guard valuable development products. Mature equipment and programming organizations, with single-digit or declining development, need to get unsurprising incomes; what's more, maintain faith in plan of action advancements, such as the transition to repeated income.

While artificial intelligence (AI) has been present since the 1950s, it has only recently become affordable and clever enough to be employed in business systems. In fact, this is a subset of a larger group of Artificial Intelligence innovations focused on using factual tactics to help computers "learn.", i.e., Examine data, uncover patterns, then use those examples to make predictions or decisions.

AI requires a training dataset, such as a large volume of recorded transaction exchanges, as well as introduction rules from a human administrator on how to get started. The calculation then, at that point, shows up at discoveries freely and, by Constantly renewing its reproductions, improves the computations for enhancing the comparability of specific traits. The computer "learns" from the data, collects more data to improve, and gets better at predicting which behaviors produce the best results. The bigger the expectations, the more preparation information and learning cycles there are; hence, AI champions benefit from both size and first-mover advantages.

Following quite a while of outstanding upgrades in handling power, distributed computing /stockpiling, and many years of corporate interests in IT frameworks, the advanced impression of most organizations currently incorporates a total history of deals associations, promoting exercises, buys, the Client administrative tickets, web-based media chatter, data from web of things (IoT) sensors, and other conditional knowledge are all used in Saas steps. Add to that an almost limitless stockpile of socioeconomic data from foreign markets, full scale monetary metrics, financial measures, climatic models, and other vital facts currently available to drive have also been completed. The chances for AI have never been exceptional to change the way organizations run and develop their business.

REFERENCES

1. Y. Parikh and E. Abdelfattah, "Clustering Algorithms and RFM Analysis Performed on Retail Transactions," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020, pp. 0506-0511, doi: 10.1109/UEMCON51285.2020.9298123.

2. S. Wu, W. -C. Yau, T. -S. Ong and S. -C. Chong, "Integrated Churn Prediction and Customer Segmentation Framework for Telco Business," in IEEE Access, vol. 9, pp. 62118-62136, 2021, doi: 10.1109/ACCESS.2021.3073776.

3. D. Bin, S. Peiji and Z. Dan, "Data Mining for Needy Students Identify Based on Improved RFM Model: A Case Study of University," 2008 International Conference on Information Management, Innovation Management and Industrial Engineering, 2008, pp. 244-247, doi: 10.1109/ICIII.2008.128.

4. A. Miklosik, M. Kuchta, N. Evans and S. Zak, "Towards the Adoption of Machine Learning-Based Analytical Tools in Digital Marketing," in IEEE Access, vol. 7, pp. 85705-85718, 2019, doi: 10.1109/ACCESS.2019.2924425.

5. Yi Wang, Qixin Chen, Chongqing Kang, Mingming Zhang, Ke Wang and Yun Zhao, "Load profiling and its application to demand response: A review," in Tsinghua Science and Technology, vol. 20, no. 2, pp. 117-129, April 2015, doi: 10.1109/TST.2015.7085625.

6. C. K. Leung, Y. Chen, S. Shang and D. Deng, "Big Data Science on COVID-19 Data," 2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE), 2020, pp. 14-21, doi: 10.1109/BigDataSE50710.2020.00010.

7. M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," in IEEE Access, vol. 7, pp. 106111-106123, 2019, doi: 10.1109/ACCESS.2019.2930410.

8. I. Khan, J. Z. Huang, M. A. Masud and Q. Jiang, "Segmentation of Factories on Electricity Consumption Behaviors Using Load Profile Data," in IEEE Access, vol. 4, pp. 8394-8406, 2016, doi: 10.1109/ACCESS.2016.2619898.

9. E. Umuhoza, D. Ntirushwamaboko, J. Awuah and B. Birir, "Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa," in SAIEE Africa Research Journal, vol. 111, no. 3, pp. 95-101, Sept. 2020, doi: 10.23919/SAIEE.2020.914260.