2023

DATA EHTICS

PIEUSH VYAS
(UOA-MS-DS)

Table of Contents

1. ABOUT THE DATASET

The dataset provided for analysis contains information related to customer orders and their associated details. Based on the dataset provided, here is a brief description of each column:

ATTRIBUTE NAME	DESCRIPTION
Order	A unique identifier for each order
Member	A unique identifier for each member/customer.
SKU	A list of product identifiers associated with each order.
Created	The date when the order was created.
Description	Description of the products included in the order.
Member's Full Name	Full name of the member/customer.
Member's Address	Address of the member/customer.
Member's Email	Email address of the member/customer.
Member's Phone Number	Phone number of the member/customer.
Member's Gender	Gender of the member/customer (Male or Female).
Member's Date of Birth	Date of birth of the member/customer.
Member's Membership Level	Membership level of the member/customer (e.g., Gold, Silver, and Bronze).
Member's Purchase History	Number of previous purchases made by the member/customer.
Order Value	Total value of the order.
Payment Method	Payment method used for the order (e.g., PayPal, Credit Card, Cash).
Delivery Address	Address to which the order is delivered.
Order Status	Status of the order (e.g., Shipped, Delivered, and Pending).
Credit Card Number	Credit card number used for payment (masked in the provided dataset).

The dataset contains **8387** records with **18** attributes which will be used for predicting customer churn risk. These attributes include sensitive information (such as personal identification details) and non-sensitive information (such as order details). The goal is to ensure the protection of sensitive information while still allowing for meaningful analysis of non-sensitive attributes to derive insights for business decisions. This involves applying techniques like data classification, minimization, masking, and anonymization to balance privacy.

2. PROBLEM STATEMENT

As an online grocery delivery company, “Everyday Deals,” you are part of the data science team to scale up the business with the power of data. As part of the organization’s attempts to increase its sales, revenue, and profit and ultimately cover more markets in the country, your team is tasked to generate better product recommendations to existing customers so that there is higher customer engagement and reduced churn rate.

3. DATA CLASSIFICATION

Data Classification involves identifying and categorizing the various types of data in the dataset based on their sensitivity and importance. It helps in ensuring that appropriate privacy measures are applied to sensitive information. In the context of the dataset provided, data classification would involve:

- **Personally Identifiable Information (PII):** Identifying attributes that directly relate to individual identity, such as names, addresses, phone numbers, email addresses, and credit card numbers. These attributes are considered highly sensitive and require special protection.
- **Sensitive Attributes**: Apart from PII, there might be other attributes that are sensitive in nature, such as birthdates or membership levels, which can still reveal information about the individual. These attributes need to be identified and treated with care.
- **Non-Sensitive Attributes**: Identifying attributes that do not contain personal or sensitive information, such as order details, SKUs, order values, and order status. These attributes are less sensitive and are analysed more freely.
- **Categorical vs. Numerical Data**: Classifying attributes as categorical (qualitative) or numerical (quantitative) help in understanding the nature of the data and applying appropriate privacy measures.

Sensitive information within the dataset, such as personally identifiable information (PII), has been carefully identified and classified. This involves recognizing attributes like names, addresses, phone numbers, email addresses, and credit card numbers that can potentially lead to the identification of individuals. The goal is to segregate and treat these attributes differently to ensure their proper protection. Sensitive and non-sensitive data features are classified and handled accordingly. As per the dataset sensitive features are Member’s Name, Member’s Address, Member’s Phone number, Member’s Email and Credit card detail.

4. DATA MINIMIZATION

Data Minimization is a principle in data privacy that involves reducing the amount of personal and sensitive data which is collected, processed, and then stored to the minimum necessary for achieving the intended purpose. The goal is to limit the exposure of sensitive information, thereby reducing the risk of data breaches and unauthorized access. In the context of the provided dataset, data minimization can be entailed by:

- **Removing Unnecessary Attributes**: Identifying and removing attributes that are not essential for analysis or business operations. For example, if certain attributes like "Member's Full Name" or "Credit Card Number" are not required for analysis, they are removed from the dataset.
- **Storing Sensitive Data Separately**: If certain sensitive attributes are not needed for immediate analysis but are required for specific tasks (e.g., customer support), these attributes can be stored separately with restricted access. This limits the exposure of sensitive information to only authorized personnel. Here in this dataset it was not required.
- **Aggregation**: Instead of storing individual-level data, aggregating data can help in preserving privacy. For instance, rather than storing each purchase individually, total purchase history for a member can be stored.

We have done data minimization by removing unnecessary columns which carry the sensitive information. This involves removing attributes like names, addresses, phone numbers, email addresses, and credit card numbers that can potentially cause data harm leading to the identification of individuals. Our goal is to minimize the attributes to ensure their proper protection. As a result organizations collect and retain only the information that is truly necessary, reducing the potential impact of data breaches and enhancing overall data security.

5. ATTRIBUTE MASKING

Attribute masking is a data privacy technique used to protect sensitive information by partially obscuring or hiding certain attributes within a dataset. This is done to ensure that while the data remains usable for analysis or processing, the sensitive aspects are not fully exposed. Here's how attribute masking works and why it's important:

- **Partial Concealment**: Attribute masking involves selectively masking or obfuscating certain characters or portions of the data. For example, in the context of member phone numbers, only a few digits might be shown, and the rest are replaced with asterisks. This prevents full exposure of sensitive information while retaining enough details for analysis.
- **Security Enhancement**: Masked attributes help prevent unauthorized access or misuse of sensitive data. Even if someone gains access to the data, they will not have full access to critical details like complete credit card numbers, addresses, or phone numbers.
- **Regulatory Compliance**: Many data privacy regulations, such as GDPR, require organizations to protect sensitive information. Attribute masking assists in complying with such regulations by reducing the risk of data breaches or leaks.

In provided dataset attributes which were required to be masked are ‘Member’s phone number’, ‘Member’s name’, ‘Member’s address’, ‘Member’s Email’ and ‘Credit card detail’. These attributes where chosen since these columns holds sensitive information about members. These attributes are masked or anonymized to protect customer identities while still preserving the integrity of the dataset.

6. NON-SENSITIVE ATTRIBUTE IDENTIFICATION

There are multiple attributes that are not sensitive, yet are valuable for analysis and are identified. These attributes provide insights that contribute to decision-making and understanding customer behaviour. Non sensitive and valuable attributes present in our dataset are 'Member ID', 'Age' ,'Payment Method' which are crucial for deriving meaningful insights. These attributes contribute to understanding customer behaviour, preferences, and trends, enabling informed decision-making and targeted strategies for enhancing customer experience and operational efficiency. Here's a detailed explanation of this process:

- **Value for Analysis**: Non-sensitive attributes often contain valuable information for analysis. For example, 'Member ID' is used to track individual shopping behaviour, 'Order Count' provided insights into customer engagement, 'Order Value' helped in understanding spending patterns, 'Age' is used for age-based segmentation, and 'Created On' helped to analyse trends over time.
- **Correlations:** Investigating correlations between non-sensitive attributes and other variables in the dataset. For instance, we find that 'Order Count' has a positive correlation with customer loyalty.
- **Business Strategy**: Non-sensitive attributes can influence business strategies. For example, insights gained from analysing 'Order Count' might lead to loyalty programs for customers with a high count, and 'Age' might guide marketing campaigns targeting specific age groups.
- **Machine Learning Features**: Non-sensitive attributes is commonly used as features in machine learning models. For instance, in this churn prediction model, 'Order Count,' 'Order Value,' and 'Age' was a crucial features that influenced the prediction.

In conclusion, non-sensitive attributes provides deeper understanding of customer behaviour and play a crucial role in various business processes and strategies by identifying hidden trends and pattern in the dataset.

7. NON-PERSONAL ATTRIBUTE IDENTIFICATION

Attributes that are neither sensitive nor personally identifiable in nature are recognized. These attributes provided valuable insights when aggregated and analysed. Non-personal attributes included Description, Order count, Order Value, Created On and purchase quantities. Leveraging these attributes allows for meaningful analysis without violating individual privacy. Analysing the shopping pattern of customers provided valuable insights into their likelihood to churn or not. Here's a breakdown of how this analysis works:

- **Shopping Behaviour**: By examining the historical data of customers' shopping activities, such as the frequency of orders, order values, and total order count we identified patterns. For instance, a customer who frequently places orders and spends a significant amount for consecutive 2 years i.e. 2013 & 2014, indicates loyalty and a lower churn risk. Similarly a customer who don't frequently places orders and spends a low amount indicates a high churn risk.
- **Churn Analysis**: Once segmented, we analysed the churn rates for each category. We found that "high-value" customers have a lower churn rate compared to other segments, it indicates that customers who spend more are more likely to remain loyal. On the other hand, if customers have a high churn rate, it implies that those who only made a less purchase are more likely to leave.
- **Predictive Modelling**: We build predictive models using machine learning techniques to forecast churn. By using features like order frequency, total order value and total order count, the model learn to predict whether a customer is likely to churn in the future. We have implied various machine learning algorithms such as 'Logistic Regression', 'Decision Tree classifier', 'Random Forest Classifier', 'Gradient Boosting', 'Hyper parameter tuning with random forestCV' and 'Voting classifier' to analyse our data. Among all these models the 'Decision tree', 'Random Forest Classifier', 'Gradient Boosting', and 'Voting classifier' showed impressive results in predicting churned and retained customers.

We can conclude that the analysing shopping pattern helps in understanding customer behaviour and predicting their likelihood to churn. This knowledge enables businesses to implement proactive measures to retain customers and tailor marketing efforts effectively.

8. DATA ANONYMIZATION

Data anonymization is a process that involves altering or transforming data in a way that individual identities are protected while still allowing meaningful analysis to be performed. The goal is to achieve a balance between data usability and privacy protection, ensuring that sensitive information cannot be easily linked back to specific individuals.

8.1. IMPORTANCE OF DATA ANONYMIZATION

Following points describes the importance of data anonymization.

- **Preserving Privacy**: Anonymization helps in safeguarding individuals' sensitive information, preventing unauthorized access and breaches.
- **Regulatory Compliance**: Many data protection regulations, such as GDPR, require organizations to ensure the privacy of individuals' data. Anonymization helps meet these legal requirements.
- **Data Sharing**: Organizations can share anonymized data with researchers or partners without risking the exposure of personal information.
- **Ethical Considerations**: Anonymization addresses ethical concerns by minimizing the potential harm that can arise from data misuse.

8.2. STEPS FOR DATA ANONYMIZATION

Data anonymization basically involves two steps. Encrypting and decrypting the sensitive data are the two major steps.

8.2.1 ENCRYPTING THE SENSITIVE DATA

There are multiple techniques of doing Data Anonymization. Few of them are:

- **De-identification:** De-identification is the process of removing or altering information from the dataset in order to protect individual privacy. It involves transforming the data in a way that it can no longer be linked back to specific individuals. In this step we delete all the sensitive information from our

dataset and create a new data frame which can we further used for data processing and making decision. Sensitive attributes such as Member's Name, Member's Address, Member's Phone number, Member's Email and Credit card detail are dropped and then a new data frame is created.

- **Data Masking**: Data masking, also known as data obfuscation or data anonymization, is a technique used to protect sensitive information by replacing, hiding, or scrambling original data with fake or altered data while maintaining the data's overall format and structure. The primary goal of data masking is to ensure that sensitive data remains confidential and secure, especially when it needs to be shared or used in non-production environments for testing, development, or analysis. In this dataset data masking is applied on the 'Member Phone Number' attribute. As a result the phone number got partially masked which means few of the digits of phone number got replaced with a special character and few of the digits remain the same. In this way we were successful in hiding the sensitive information along with maintaining the integrity of the dataset.

- **Pseudonymization**: The importance of encrypted data lies in the protection and security it provides to sensitive information. Encryption is a crucial tool in safeguarding data from unauthorized access, ensuring privacy, and preventing data breaches. Here are some key reasons why encrypted data is important:

    a. Data Confidentiality
    b. Data Integrity
    c. Privacy Protection
    d. Mitigating Data Breaches
    e. Secure Communication
    f. Compliance and Legal Requirements
    g. Protecting Intellectual Property
    h. Cloud Security

To encrypt the specified columns by using Fernet encryption, the cryptography library in Python is used. With the help of this library all the sensitive data has been encrypted. Now if this encrypted data is shared with anyone then they will not be able to misuse any sensitive information since all the sensitive information has been encrypted.

8.2.2 DECRYPTING SENSITIVE DATA

Decryption is the process of reversing encryption. If the dataset has been encrypted to protect sensitive information, decryption is required to retrieve the original data. Encryption transforms data into a secure format using an encryption algorithm and a key. Only individuals with the decryption key can transform the data back to its original form. Decrypting the data should be done with proper authorization and security measures in place.

9. CONCLUSION

In summary, all the process which are described involves first de-identifying the data through techniques like data masking and Pseudonymization to protect sensitive information while maintaining the usability of the dataset. If encryption has been used, decryption with the appropriate key is required to retrieve the original data. These steps are crucial for ensuring data privacy and complying with regulations while still allowing for analysis and insights.

Moreover based on the analysis and evaluation of different classification models, here are the conclusions:

- **Logistic Regression**: The logistic regression model achieved an accuracy of 67%. It performed well in predicting class 1 (churned customers) with high precision, recall, and f1-score. However, it struggled with class 0 (Not churned customers) and showed lower precision, recall, and f1-score for this class.
- **Random Forest**: The random forest model achieved an accuracy of 81%. It performed well in predicting both classes with high precision, recall, and f1-score. The model's ensemble nature helped in reducing overfitting, and it shows promise in making accurate predictions.
- **Decision Tree**: The decision tree model achieved an accuracy of 100%. It showed perfect performance in both classes, achieving 100% precision, recall,

and f1-score for both churned and retained customers. However, there is a possibility of overfitting.

- **Hyper Parameter tuning**: The Hyper tuned model achieved an accuracy of 79%. It performed well in predicting both classes with high precision, recall, and f1-score.Overall, the classifier is performing flawlessly with high accuracy, precision, recall, and F1-score, which is a good for a model.
- **Gradient Boosting**: The gradient boosting model achieved an accuracy of 100%. Similar to the Decision Tree, it performed very well in predicting both classes with high precision, recall, and f1-score. The ensemble boosting technique improved the model's performance.
- **Soft Voting Ensemble**: The soft voting ensemble of different classifiers achieved an accuracy of 91.27%. By combining the predictions of individual models, it improved overall performance and showed competitive results compared to individual models.

In conclusion, ‘Decision tree’, ‘Gradient Boosting’, and ‘Voting classifier’ showed impressive results in predicting churned and retained customers. However, ‘Random Forest’ also showed a good result.

****