ML Project for Customer Lifetime Value Prediction in Telecom: Learning Material

Data Exploration

Data exploration is a critical step in the data analysis process, where you examine the dataset to gain a preliminary understanding of the data, detect patterns, and identify potential issues that may need further investigation. Data exploration is important because it helps to provide a solid foundation for subsequent data analysis tasks, hypothesis testing, and data visualization.

Data exploration is also important because it can help you identify an appropriate data analysis approach.

Here are the various functions that help us explore and understand the data.

- Shape: Shape is used to identify the dimensions of the dataset. It gives the number of rows and columns present in the dataset. Knowing the dimensions of the dataset is important to understand the amount of data available for analysis and to determine the feasibility of different methods of analysis.
- Head: The head function is used to display the top five rows of the dataset. It helps us to understand the structure and organization of the dataset. This function gives an idea of what data is present in the dataset, what the column headers are, and how the data is organized.
- Tail: The tail function is used to display the bottom five rows of the dataset. It provides the same information as the head function but for the bottom rows. The tail function is particularly useful when dealing with large datasets, as it can be time-consuming to scroll through all the rows.
- Describe: The describe function provides a summary of the numerical columns in the dataset. It includes the count, mean, standard deviation, minimum, and maximum values, as well as the quartiles. It helps to understand the distribution of the data, the presence of any outliers, and potential issues that can affect the model's accuracy.
- Isnull: The isnull function is used to identify missing values in the dataset. It returns a Boolean value for each cell, indicating whether it is null or not. This

function is useful to identify the presence of missing data, which can be problematic for regression analysis.

- Dropna: The dropna function is used to remove rows or columns with missing data. It is used to remove any observations or variables with missing data, which can lead to biased results in the regression analysis. The dropna function is used after identifying the missing data with the isnull function.
- Columns: The .columns method is a built-in function that is used to display the column names of a pandas DataFrame or Series. It returns an array-like object that contains the names of the columns in the order in which they appear in the original DataFrame or Series. It can be used to obtain a quick overview of the variables in a dataset and their names.

Outlier Detection:

Outlier detection is a critical data analysis technique that involves identifying and removing data points that are significantly different from the rest of the data. Outliers are data points that lie far away from the rest of the data, and they can significantly influence the statistical analysis and machine learning models' performance. Therefore, identifying and removing outliers is essential to ensure accurate and reliable data analysis results.

There are two main approaches for outlier detection: parametric and non-parametric.

- Parametric Methods: Parametric methods assume that the data follows a specific distribution, such as a normal distribution. In this approach, outliers are identified by calculating the distance of each data point from the mean of the distribution in terms of the number of standard deviations. Data points that are beyond a certain number of standard deviations (usually three or more) are considered as outliers.

One common parametric method is the Z-score method, which calculates the distance of each data point from the mean in terms of standard deviations. Parametric methods can be useful when the data follows a known distribution, but they may not be effective when the data is not normally distributed.

- Non-Parametric Methods: Non-parametric methods do not assume any specific distribution of the data. Instead, they rely on the rank or order of the data points. In this approach, outliers are identified by comparing the values of each data point with the values of other data points. Data points that are significantly different from other data points are considered as outliers.

Quantiles are an important concept in non-parametric outlier detection methods. They represent values that divide a dataset into equal-sized parts, such as quarters or thirds. The most commonly used quantiles are the median (which divides the data into two equal parts), the first quartile (which divides the data into the lowest 25% and the highest 75%), and the third quartile (which divides the data into the lowest 75% and the highest 25%).

The interquartile range (IQR) is another important concept related to quantiles. It is defined as the difference between the third and first quartiles and represents the middle 50% of the data. The IQR can be used to identify outliers by defining a range (known as the Tukey's fence) beyond which any data points are considered outliers.
Non-parametric methods can be useful when the data is not normally distributed or when the distribution is unknown.

By calculating quantiles for each continuous variable in the dataset, we are trying to get an idea about the spread and distribution of the data. Specifically, we are interested in identifying potential outliers in the data.

Quantiles divide a distribution into equal proportions. For instance, the 0.25 quantile is the value below which 25% of the observations fall and the 0.75 quantile is the value below which 75% of the observations fall. By calculating quantiles at various levels, we can get a better understanding of the distribution of the data and identify any observations that are too far away from the rest of the data.

These quantiles can be used as thresholds to identify potential outliers in the data. Observations with values beyond these thresholds can be considered as potential outliers and further investigation can be carried out to determine if they are true outliers or not.

Assumptions:

To later calculate customer lifetime value (CLTV), we are making several assumptions to simplify the calculation process.

- Incoming usage is free: We assume that customers are not charged for incoming calls or messages. This assumption is made because it is difficult to track and measure incoming usage, and because incoming usage is often not a significant source of revenue for telecom companies.

- Outgoing to call center is free: We assume that customers are not charged for calls made to the call center. This assumption is made because calls to the call center are typically made to address customer service issues, and charging customers for these calls may discourage them from seeking help when they need it.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the process of analyzing and summarizing data to find patterns, relationships, and insights that can help inform business decisions. EDA is an important step in data analysis as it allows us to understand the data and identify potential problems such as outliers, missing values, or data inconsistencies. EDA can involve visualizing data using graphs and charts, calculating summary statistics such as means and medians, and identifying correlations between variables.

Correlation

Correlation coefficient is used to measure the strength of relationship between two variables. It indicates that as the value of one variable changes the other variable changes in a specific direction with some magnitude. There are various ways to find correlation between two variables, one of which is Pearson correlation coefficient. It measures the linear relationship between two continuous variables.

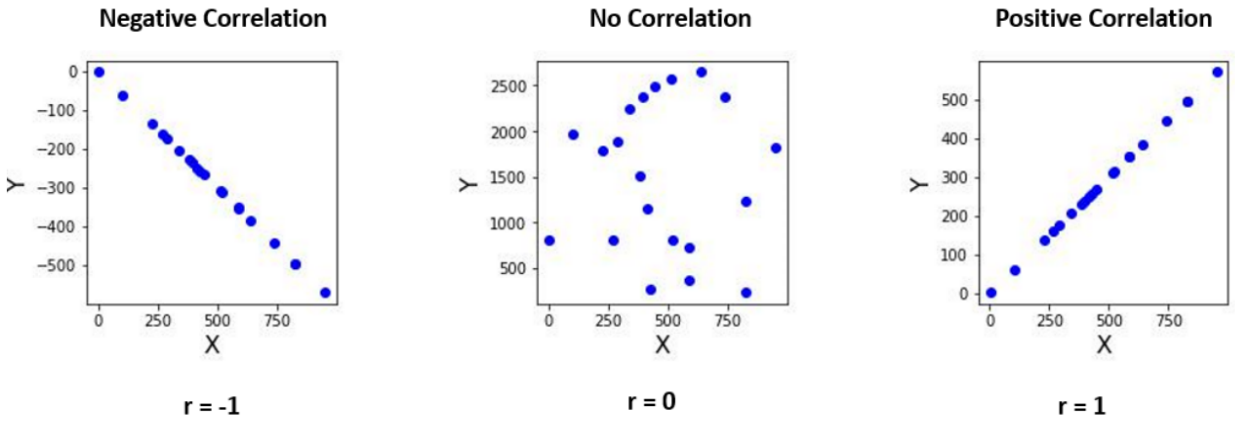
Let's say x and y are two continuous variables, the Pearson correlation coefficient between them can be found by the following formula.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i and y_i represents the i^{th} value of the variables. The value of r ranges between -1 and $+1$.

Their strength of relationship is measured by the absolute value of coefficient, whereas the sign of the coefficient indicates the direction of the relationship.

Graphs of Different Correlation Coefficients



1. $r = -1$ indicates a perfect negative relationship between the variables
2. $r = 0$ indicates no relationship between the variables
3. $r = 1$ indicates a perfect positive relationship between the variables

T-Tests

A t-test is a statistical hypothesis test that compares the means of two groups. It is commonly used to determine if there is a significant difference between the means of two groups, such as a control group and a treatment group. The test is based on the t-distribution, which is a probability distribution that describes how the means of random samples from a population with a normal distribution are distributed.

There are two types of t-tests: the one-sample t-test and the two-sample t-test. The one-sample t-test is used to compare the mean of a sample to a known value, while the two-sample t-test is used to compare the means of two independent samples.

To perform a t-test, the first step is to formulate the null and alternative hypotheses. The null hypothesis is that there is no significant difference between the means of the two groups, while the alternative hypothesis is that there is a significant difference between the means of the two groups.

The next step is to calculate the test statistic, which is the t-value. This is done by taking the difference between the means of the two groups and dividing it by the standard error of the difference between the means. The standard error is a measure of the variability of the sample means and is calculated using the sample size and the standard deviation of each group.

Finally, the t-value is compared to a critical value from the t-distribution based on the degrees of freedom and the significance level chosen. If the calculated t-value is greater than the critical value, the null hypothesis is rejected, and it is concluded that there is a significant difference between the means of the two groups.

Pair Plots

Pair plots are a type of visualization that displays the pairwise relationships between variables in a dataset. They are particularly useful in identifying patterns and relationships between multiple variables at once. Pair plots are essentially a grid of scatterplots that show the relationships between each pair of variables in a dataset.

Each variable is plotted against every other variable, so the diagonal of the plot shows the distribution of each variable. The off-diagonal plots show the scatterplot of the two variables.

Pair plots are helpful in identifying correlations between variables, patterns in the data, and any potential outliers or anomalies. They can also help in identifying which variables are most strongly correlated with the target variable, which can be useful in predictive modeling or feature selection.

Customer Lifetime Value

Customer Lifetime Value (CLTV) is a key metric used to measure the total monetary value a customer brings to a business over the entire period of their relationship with the business. In general, there are several ways to calculate CLTV, including revenue-based, profit-based, and customer-based approaches. The revenue-based approach is the most commonly used, and it considers factors such as average revenue per user (ARPU), service cost, and customer lifetime. The profit-based approach takes into account the net profit generated by a customer, while the customer-based approach looks at the value of a customer based on factors such as customer acquisition cost and retention rate.

Overview of different CLTV Calculation Methods

- **Historical Method**: This method calculates CLTV based on the past behavior of customers. It involves analyzing the transactional history of customers, including average purchase value, and length of the relationship. The historical method assumes that the customer behavior will remain the same in the future.
- **Predictive CLTV**: This method involves using predictive modeling techniques to estimate the future value of customers. Predictive CLTV takes into account various factors such as customer demographics, past behavior, and purchase history. Machine learning algorithms such as regression analysis, decision trees, and neural networks can be used to build predictive models for CLTV. The data we currently are using is not suitable for predictive CLTV because we don't have a labeled CLTV column, and we are calculating it based on other features such as ARPU, service costs, and churn rates. Predictive CLTV requires labeled data, and we need to build a predictive model that can accurately predict the CLTV based on the available features. This would require collecting more data on customer behavior and transactions, which can be used to train a machine learning model to predict CLTV.
- **Probabilistic CLTV**: This method uses probability theory to estimate the value of a customer. It takes into account the uncertainty in the future behavior of customers and assigns probabilities to different outcomes. This method is useful when there is a high degree of uncertainty in customer behavior.
- **Cohort-based CLTV**: In this approach, customers are grouped based on when they started doing business with the company (e.g., monthly or quarterly cohorts). This allows for a comparison of how different cohorts behave over time, which can inform CLTV calculations. Cohort-based CLTV can provide easier comparisons over the group of people.
- **Customer Segmentation CLTV**: This approach groups customers into different segments based on their behavior and demographics, and calculates CLTV separately for each segment. This allows companies to identify the most valuable customer segments and develop tailored marketing strategies.

CLTV Calculation: Historical Data

To calculate the revenue-based CLTV for all customers, we need to use historical data to estimate the average revenue per user (ARPU), service cost, and customer lifetime. Once we have these metrics, we can use the following formula to calculate CLTV:

$$\text{Historical CLTV} = \frac{\text{ARPU} - \text{Average service cost per user}}{\text{Churn rate}}$$

The average service cost per user is the average cost of providing services to each user, such as marketing, customer support, and infrastructure costs. The churn rate is the percentage of customers who stop using the company's services over a given period.

By using this formula, we can estimate the total value that a customer will generate over their entire relationship with the company based on historical data.

Assumption:

We are assuming a profit margin of average 10%.

Derivation

Since we assume a 10% profit margin, the average service cost per user is equal to 90% of the ARPU:

$$\text{Average service cost per user} = 0.9 \times \text{ARPU}$$

Substituting the above equation in the formula for historical CLTV, we get:

$$\text{Historical CLTV} = \frac{\text{ARPU} - 0.9 \times \text{ARPU}}{\text{Churn rate}} = \frac{0.1 \times \text{ARPU}}{\text{Churn rate}}$$

The customer lifetime can be calculated as the reciprocal of the churn rate:

$$\text{Customer Lifetime} = \frac{1}{\text{Churn rate}}$$

The historical approach can be used to calculate CLTV and use it for different perspectives such as revenue, customer churn, loyalty, and retention. This can help companies make informed decisions on how to allocate resources and target specific customer segments to maximize their profits and minimize their churn rate.

Rule Based Customer Segmentation

Segmenting customers is important while calculating Customer Lifetime Value (CLTV) because different customer segments have different behaviors and values. By segmenting customers, we can identify the most valuable customers and allocate resources to them accordingly, while also developing targeted marketing strategies to retain and increase the value of less valuable customers.

For example, high-value customers who make frequent purchases and are loyal to the brand may require personalized attention, incentives, and rewards to maintain their loyalty and increase their lifetime value. On the other hand, low-value customers who make infrequent purchases or are at risk of churning may require targeted campaigns and promotions to re-engage them and improve their value.

Segmentation can also help identify opportunities for cross-selling and upselling by analyzing customer behavior, preferences, and purchase history. By understanding the needs and wants of different customer segments, businesses can develop tailored

product offerings and marketing messages to increase customer satisfaction and drive revenue growth.

Rule-based customer segmentation is one approach to segmentation that involves dividing customers based on specific rules or criteria.

The segmentation can be based on factors such as total recharge, outgoing call usage, data usage, and satisfaction scores. By categorizing customers based on these factors, we can identify high-value customers, at-risk customers, and customers who are likely to churn. This information can then be used to develop personalized marketing campaigns and retention strategies.

Dynamic Churn Rates

Monthly dynamic churn rate is a metric that is widely used in the telecom industry to measure customer attrition. It refers to the percentage of customers who discontinue their service with a telecom provider during a given month. The dynamic nature of the churn rate reflects the constant influx of new customers as well as the loss of existing ones, and is influenced by various factors such as pricing, service quality, network coverage, customer support, and competitive offerings.

For telecom companies, managing churn rate is crucial as it directly impacts revenue and profitability. High churn rates can lead to significant revenue loss as well as increased costs associated with customer acquisition and retention efforts. On the other hand, low churn rates can result in higher customer lifetime value, increased revenue per user, and improved brand loyalty.

Usage Patterns and Deviations

Usage patterns and monthly deviations are important metrics for understanding customer behavior in the telecom industry. In this project, we focus on outgoing call usage and data usage as the main usage patterns as they are revenue drivers. To calculate monthly deviations, we subtract the previous month's usage from the current month's usage.

Before calculating deviations, we remove unnecessary columns such as incoming usage since we assume it is free, and we also remove outgoing usage to the call center since it is assumed to be free. We then calculate the deviations using the .diff() method which computes the difference between consecutive rows.

Using these usage patterns and monthly deviations, we can gain insights into customer behavior such as which customers are heavy users of outgoing calls or data, and which customers have significant fluctuations in usage from month to month.

Customer Segments

The code provided in the next section categorizes customers into four groups: new customers, churned customers, loyal customers, and retained customers. Each group is determined based on a set of rules that take into account a variety of factors such as total recharge amount, outgoing call and data usage, and satisfaction score.

Loyalty Score = (Total Recharge Deviation * w1) + (Usage Deviation * w2) + (Satisfaction Deviation * w3)

Retention Score = (Total Recharge Deviation * w4) + (Usage Deviation * w5) + (Satisfaction Deviation * w6)

Here, w1, w2, w3, w4, w5 and w6 are the weights assigned to each parameter for each segment. You can adjust the weights as per your business requirements and the impact of each parameter on the loyalty and retention of the customers in each segment.

New Customers

New customers are those who have not completed six months with the company. These are customers who have recently joined the company and may not have a strong attachment to the brand yet. Since they have no prior history with the company, their retention is likely to be more heavily influenced by their satisfaction with the product or service. Thus, higher weight could be given to the satisfaction deviation.

Churned Customers

Churned customers are those who have left the company within six months. They are given a loyalty of zero but we will calculate retention scores of these customers giving higher weight to customer satisfaction deviation.

Loyal Customers

Loyal customers are those who have completed six months with the company. These are customers who have a strong emotional attachment to the brand and are less likely to switch to a competitor. For this segment, both total recharge deviation is the most important while usage deviation and satisfaction deviation is less important.

Retained Customers

Retained customers are those who have completed six months with the company. These are customers who have been with the company for a while and have a high level of engagement. For this segment, the importance of total recharge deviation and satisfaction deviation is high, while usage deviation is relatively less important.