Customer Lifetime Value Prediction

Bhuvan Chandra, Rahul Shetty, Shruti Wakchoure

Abstract

Businesses are constantly fighting for the loyalty and retention of consumers in today's intensely competitive retail and e-commerce environment. The notion of Customer Lifetime Value (CLTV) is fundamental to this undertaking, as it is a critical statistic for forecasting the revenue a customer is anticipated to produce throughout their association with the organization. Businesses can use CLTV as a guiding concept to help them manage resources and customize their strategy in order to optimize customer value and long-term profitability.Our project uses datasets from many sources to conduct a thorough investigation of CLTV prediction. Starting with a customer churn dataset that we scraped from a GitHub repository, we complement our research with web scraped unemployment rate data and demographic information collected from the US Census Bureau's API. We can now confidently tackle important questions like estimating the lifetime value of new clients, detecting possible churners, and analyzing CLTV patterns for different customer segments and demographics thanks to this comprehensive dataset.Our goal is to deliver practical insights to organizations so they may enhance their marketing, sales, and customer service strategies through our research. Through the exploration of numerous issues related to CLTV dynamics, customer retention, and profitability, our project aims to enable companies to prosper in the dynamic e-commerce environment, promoting long-term expansion and competitiveness.

Data Collection

In the initial phase of our data collection process, we sourced a dataset from a GitHub repository focusing on customer churn analysis within the telecommunications industry. This dataset, obtained from the repository, contains a diverse range of attributes including customer demographics, satisfaction scores, churn indicators, and Customer Lifetime Value (CLTV) indices. Serving as a foundational element for our customer lifetime value prediction project, this dataset offers valuable insights into customer behaviour and retention patterns.

In the second phase of data collection, we utilized the United States Census Bureau's API to gather demographic data essential for our research. Through Python's requests package, we accessed population statistics primarily focusing on Californian cities. By carefully specifying the required parameters in the API queries, including the requested fields and geographic boundaries, we obtained detailed demographic information crucial for comprehending consumer behaviour. This meticulous approach allowed us to procure population estimates for each city, providing insights into the potential customer base across various regions.

Expanding our data collection efforts, we sought to enhance the dataset with more precise demographic statistics at the ZIP code level. Employing a systematic method, we conducted repeated queries against the Census API to gather population data for various Californian ZIP codes. This methodical approach yielded extensive demographic insights, enabling sophisticated analysis of consumer behavior across diverse geographical locations.

In the third phase of our data collection process, we employed web scraping techniques to extract unemployment rate comparisons across California ZIP codes from ZipAtlas. Commencing with regions exhibiting the highest unemployment rates, our method systematically traversed subsequent pages to compile an extensive dataset. Our aim in amalgamating these diverse data sources is to discern subtle trends and correlations, guiding strategic decisions aimed at maximizing customer lifetime value and ensuring long-term business success.

Dataset Description

The dataset includes detailed information on every column, providing an in-depth explanation of every feature and its importance in the examination of customer turnover patterns in the telecom industry.

CustomerID A unique ID that identifies each customer.

Count A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Gender The customer's gender: Male, Female

Age The customer's current age, in years, at the time the fiscal quarter ended.

Senior Citizen Indicates if the customer is 65 or older: Yes, No.

Married Indicates if the customer is married: Yes, No

Dependents Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.

Number of Dependents Indicates the number of dependents that live with the customer.

Country The country of the customer's primary residence.

State The state of the customer's primary residence.

City The city of the customer's primary residence.

Zip Code The zip code of the customer's primary residence.

Lat Long The combined latitude and longitude of the customer's primary residence.

Latitude The latitude of the customer's primary residence.

Longitude The longitude of the customer's primary residence.

Population A current population estimate for the entire Zip Code area.

Quarter The fiscal quarter that the data has been derived from (e.g. Q3).

Referred a Friend Indicates if the customer has ever referred a friend or family member to this company: Yes, No.

Number of Referrals Indicates the number of referrals to date that the customer has made.

Tenure in Months Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.

Offer Identifies the last marketing offer that the customer accepted, if applicable. Values include None, Offer A, Offer B, Offer C, Offer D, and Offer E.

Phone Service Indicates if the customer subscribes to home phone service with the company: Yes, No.

Avg Monthly Long Distance Charges Indicates the customer's average long distance charges, calculated to the end of the quarter specified above.

Multiple Lines Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No.

Internet Service Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.

Avg Monthly GB Download Indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above.

Online Security Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No.

Online Backup Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No.

Device Protection Plan Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No.

Premium Tech Support Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No.

Streaming TV Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Movies Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Music Indicates if the customer uses their Internet service to stream music from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Unlimited Data Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No.

Contract Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.

Paperless Billing Indicates if the customer has chosen paperless billing: Yes, No.

Payment Method Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check.

Monthly Charge Indicates the customer's current total monthly charge for all their services from the company.

Total Charges Indicates the customer's total charges, calculated to the end of the quarter specified above.

Total Refunds Indicates the customer's total refunds, calculated to the end of the quarter specified above.

Total Extra Data Charges Indicates the customer's total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above.

Total Long Distance Charges Indicates the customer's total charges for long distance above those specified in their plan, by the end of the quarter specified above.

Satisfaction Score A customer's overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied).

Satisfaction Score Label Indicates the text version of the score (1-5) as a text string.

Customer Status Indicates the status of the customer at the end of the quarter: Churned, Stayed, or Joined.

Churn Category A high-level category for the customer's reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price. When they leave the company, all customers are asked about their reasons for leaving. Directly related to Churn Reason.

Churn Reason A customer's specific reason for leaving the company. Directly related to Churn Category.

Unemployment Rate Unemployment rate across different zip codes.

Data Cleaning

Our main goal in the data cleaning stage was to make sure the dataset was high-quality and intact for further analysis. Several actions were taken to accomplish this. First, to maintain data accuracy and consistency, we eliminated rows that included null entries to rectify missing values. Furthermore, we found and removed unnecessary columns—such as Index, paperless billing, nation, state, Under 30, senior citizen, and latlong—that had no bearing on our study.

Additionally, outlier values were found and eliminated from the dataset to prevent them from skewing the results of our research. Statistical measurements and data distributions have to be carefully examined to find values that deviate significantly from the predicted range. We reduced the possibility of incorrect findings and guaranteed the validity of our analysis by removing outliers.

Data pretreatment and cleaning are essential processes since they have a direct effect on your models' performance. The square root and log transformations are two frequently used transformations used in this stage. These transformations are very handy when dealing with numerical data that can contain outliers or skewed distributions, which can skew statistical analysis and machine learning models. Now let's talk about applying these modifications to your pertinent numerical columns:

Many statistical and machine learning methods operate under the premise that the data is more regularly distributed, which can be achieved with great effect by using the log transformation. When data have a right-skewed distribution—that is, when the majority of the data are concentrated on the left side with a large tail on the right—it works very well.

One further method for balancing variance and reducing skewness is the square root transformation. Although it isn't as strong as the log transformation, it might be more suitable for data that is just slightly skewed or contains negative values (since the log transformation is limited to positive values).

Visualizing the altered column distributions after applying these changes is crucial for determining whether the transformation produced the intended results (e.g., reduced skewness, stabilized variance). For this, box charts or histograms might be used.Furthermore, keep in mind that to read the model's predictions in their original scale once the model has been trained, you might need to reverse the transformation (for example, by using the exponential function for log-transformed data).

In conclusion, log and square root transformations are useful techniques that can help you prepare your numerical data for more effective analysis and modeling during the data pretreatment and cleaning phase of your DM project.

Visualizations

Plot 1: Histogram of Customer Status

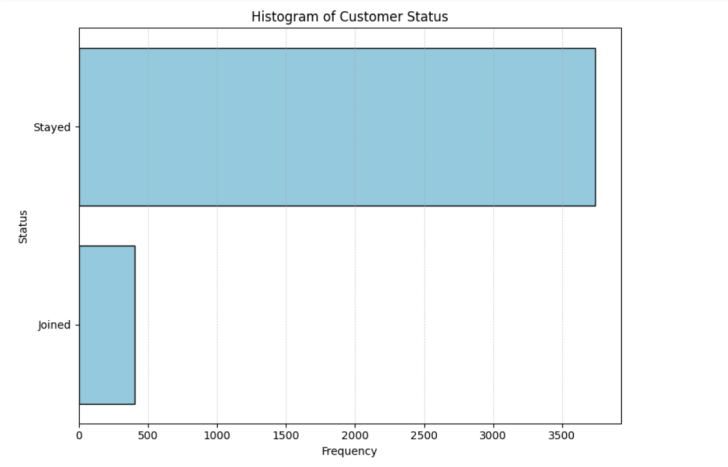

Figure 1: Histogram of Customer Status

The dataset's "Customer Status" distribution is shown by the histogram. The client statuses of "Joined" and "Stayed" are represented by two bins. The majority of "Stayed," or current, consumers are just that—customers. A smaller subset is marked as "Joined," signifying that they are fresh clients.

Plot 2: Total and Monthly Charge Distribution

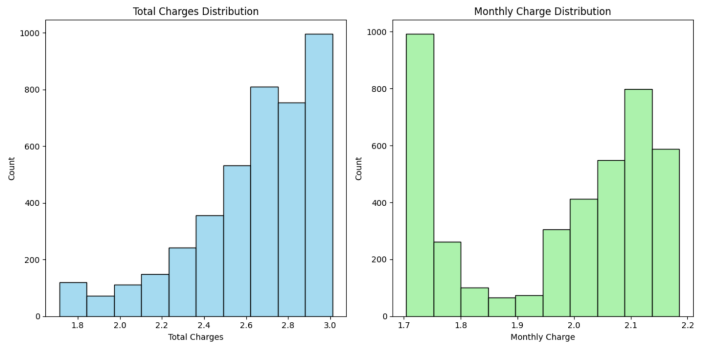

Figure 2: Total and Monthly Charge Distribution

The distribution of the total and monthly charges to clients in your dataset is shown by these histograms. The blue histogram indicates that the majority of clients have total charges that fall between the low and high ranges. According to the green histogram, more customers are charged more each month, while fewer customers are charged less. These illustrations make it easier to comprehend how costs are distributed among your customers.

Plot 3: Different charges in Total Revenue

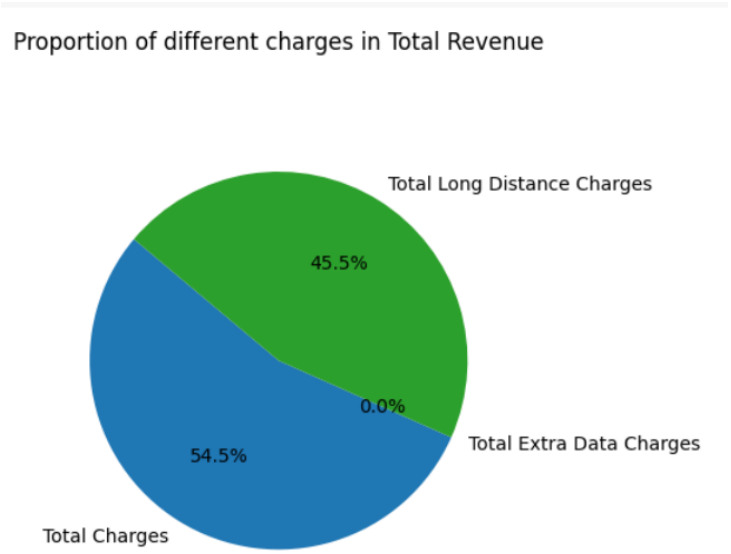

Figure 3: Different charges in Total Revenue

The pie chart shows how Total Revenue is divided into two parts: Total Charges, which constitute 54.5% of Total Revenue, and Total Long Distance Charges, which constitute 45.5%. Since their percentage is 0, the total extra data charges do not contribute to the total revenue. Understanding the proportionate contributions of the various charge kinds to the total revenue is made easier by this chart.The pie chart highlights the dominance of Total Charges and Total Long Distance Charges in the overall revenue structure and provides a clear visual depiction of the distribution of revenue components.

Plot 4: Box plots of Relevant Numerical Columns

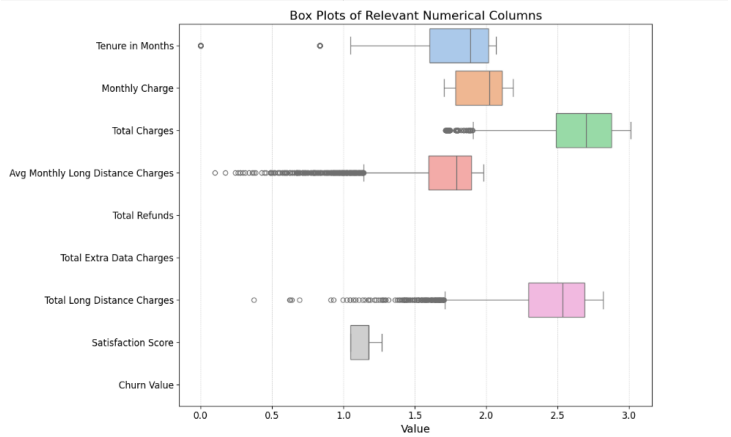

Figure 4: Box plots of Relevant Numerical Columns

- **Tenure in Months:** Displays a somewhat broad range, signifying fluctuations in the length of time clients have been associated with the business.
- **Monthly Charge:** Moreover, it has a broad IQR, indicating to clients a range of monthly rates.
- **Total Charges:** Shows a large range with a bias towards higher values.

- **Average Monthly Long Distance Fees:** While there are occasional exceptions, fees are typically modest.
- **Total Refunds:** Shows that most clients obtain refunds in a range that is narrow.
- **Total Additional Data Fees:** The tight distribution indicates that consumers' extra data fees are consistent.
- **Total Long Distance Charges:** A little bit broader range than total reimbursements or additional data fees, with multiple anomalies suggesting that certain clients have significantly higher long distance fees.
- **Satisfaction Score:** Wide range in the satisfaction score indicates different customer satisfaction levels.
- **Churn Value:** Probably a small-range or binary numeric column that indicates if a customer has churned or not.

Plot 5: Relationship between Total Charges and Monthly Charge

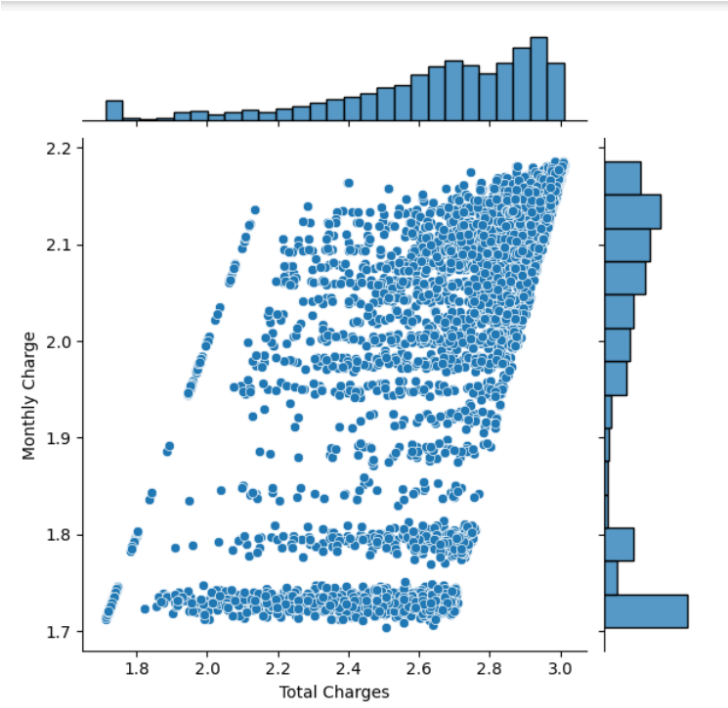

Figure 5: Relationship between Total Charges and Monthly Charge

Data points are concentrated at lower total charges and spread out as total charges rise, indicating a range of monthly charges as consumers' overall spending increases. The distribution of Total Charges is displayed in the top marginal histogram, which displays a right-skewed distribution with a concentration of consumers at the lower end. The distribution of monthly charges is depicted in the marginal histogram to the right, where it is comparatively level with a small increase toward the higher rates. This graphic aids in recognizing patterns or trends in the customer billing data and helps explain how monthly expenses affect overall charges over time.

Plot 6: Age vs Revenue

When there are too many data points to show as individual points in a scatter plot, the density of data points is represented using this image, which is a hexbin plot. The color bar on the right shows that the plot is divided into hexagonal bins, and the color of each bin indicates the number of data points within that bin. Greater concentrations of data points are represented by darker hues. Two variables,

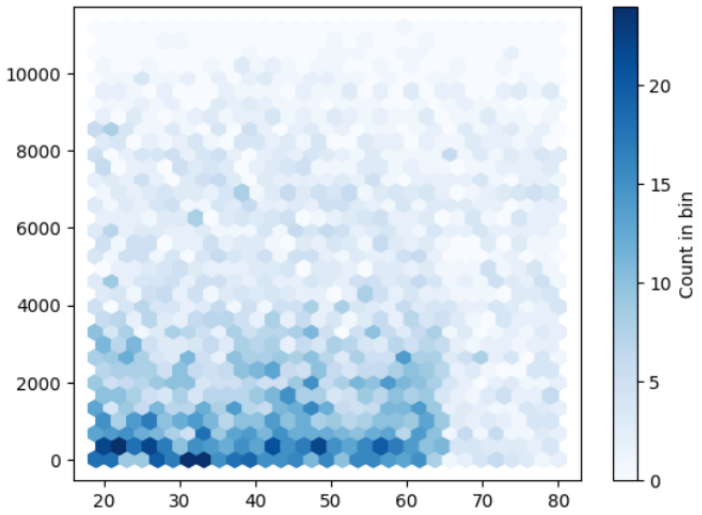

Figure 6: Age vs Revenue

one on the x- and one on the y-axes, are probably represented by the plot. Darker areas indicate a higher concentration of people in particular age and economic ranges. As an example, it might depict the relationship between age (x-axis) and income (y-axis).

Plot 7: Distribution of Total Revenue by Contract Type

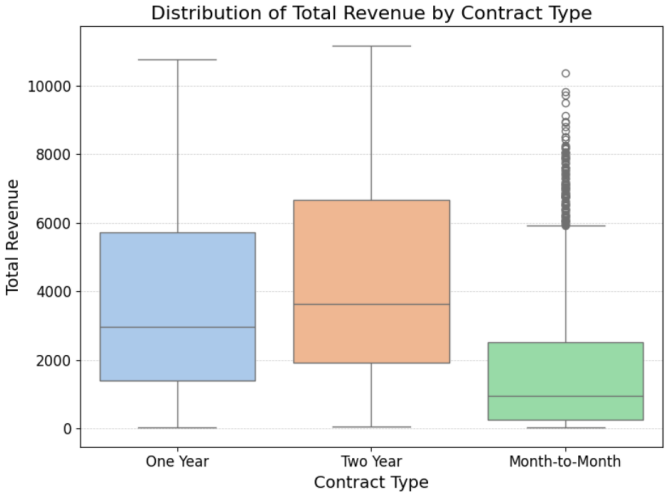

Figure 7: Distribution of Total Revenue by Contract Type

The box plot illustrates how various contract types affect Total Revenue. One-year contracts are less variable and have a median revenue in the middle range. With a wider range and numerous outliers, two-year contracts have a higher median revenue that suggests some very high revenue clients. When compared to longer contracts, month-to-month agreements had the smallest range and lowest median revenue, indicating greater stability but generally lower total revenue.

Plot 8: Survey Responses by Internet Service Status

The comparison of survey responses between consumers with and without internet connection, grouped by various satisfaction rankings, is shown in this stacked bar chart. With segments inside each bar signifying the percentage of respondents with particular satisfaction scores, each bar represents a group (No or Yes to Internet Service).

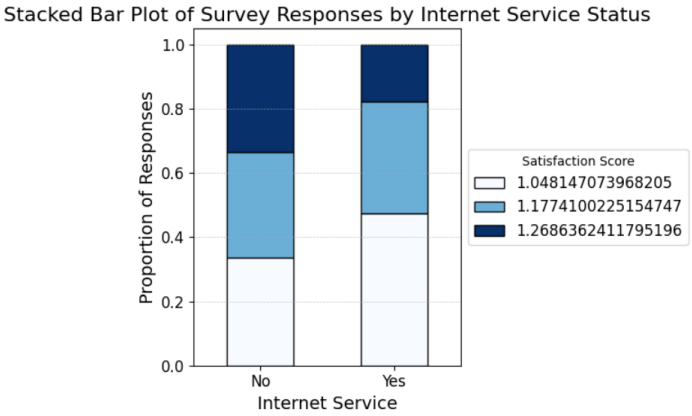

Figure 8: Survey Responses by Internet Service Status

From the lowest to the highest satisfaction scores, the different shades of blue represent the scores. A broad range of scores is shown by both groups, indicating similar score distributions. But there isn't enough labeling in the legend to understand how scores relate to color parts. Still, the graphic does a good job of illustrating how consumers' satisfaction ratings differ according to their internet service status.

Plot 9: Proportion of Customers with Internet Service

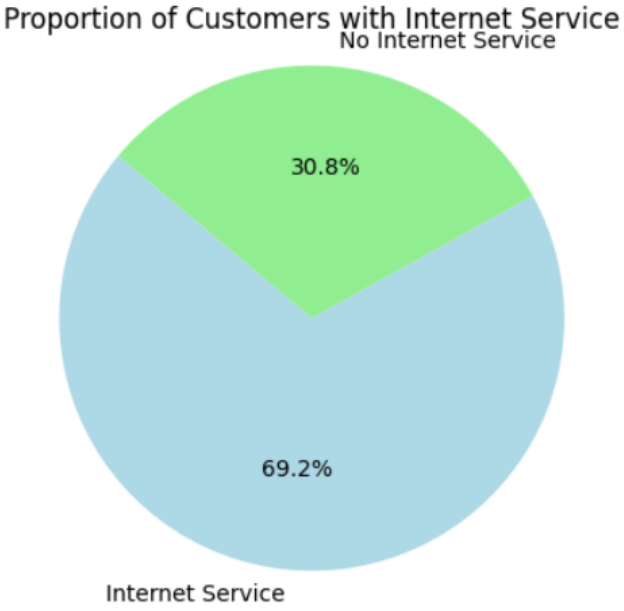

Figure 9: Proportion of Customers with Internet Service

The percentage of consumers split into two groups according to whether or not they have internet access is displayed in the pie chart. 30.8% of people do not have internet service, compared to the majority of 69.2% who do. The distribution of customers in the dataset with and without internet service is shown graphically in this figure.

Plot 10: Monthly Charge by Internet Service

The monthly costs for DSL, Fiber Optic, and Cable internet services are displayed in a scatter plot. Every dot symbolizes a customer, and its vertical position signifies the monthly fee. It shows that while Fiber Optic customers face a larger range of rates, usually greater than DSL, DSL users often have lower

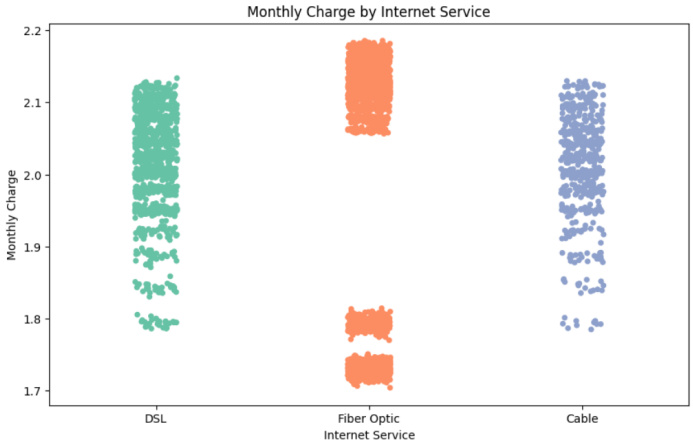

Figure 10: Monthly Charge by Internet Service

charges within a small range. The rates that cable customers pay are similar to those of DSL users, however they lean a little bit in the direction of higher rates. The plot's jitter skillfully illustrates the density of points, offering details on the concentration of users at various price tiers for every kind of internet service.

Models Implemented

Decision Tree Classifier

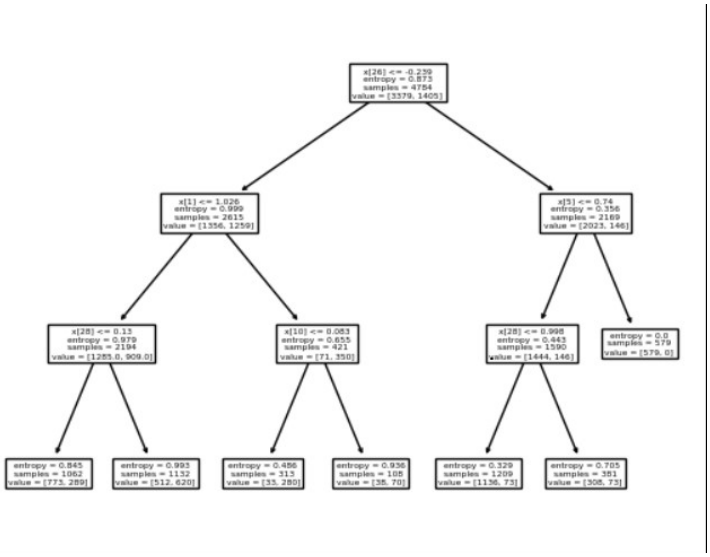

Figure 11: Decision Tree Classifier

Used to predict customer churn by learning decision rules from data features. Chosen for its interpretability and effectiveness in handling categorical data, making it suitable for business stakeholders to understand model decisions. The Decision Tree Classifier achieved a training accuracy of 78.7% and a test accuracy of 77.4%, with an F1 score of 0.632.Interpretation for Churn Prediction: The decision tree's performance suggests it can reasonably predict customer churn. Its interpretability helped in identifying key behaviors and characteristics of customers who are likely to churn.Interpretation for CLTV Prediction: By identifying the most significant factors for churn, we inferred critical aspects affecting CLTV. For instance, if 'tenure' is a key node in the tree, it implies a longer customer lifespan increases CLTV, and strategies focused on retention,to enhance overall value.

Gaussian Mixture Model (GMM)

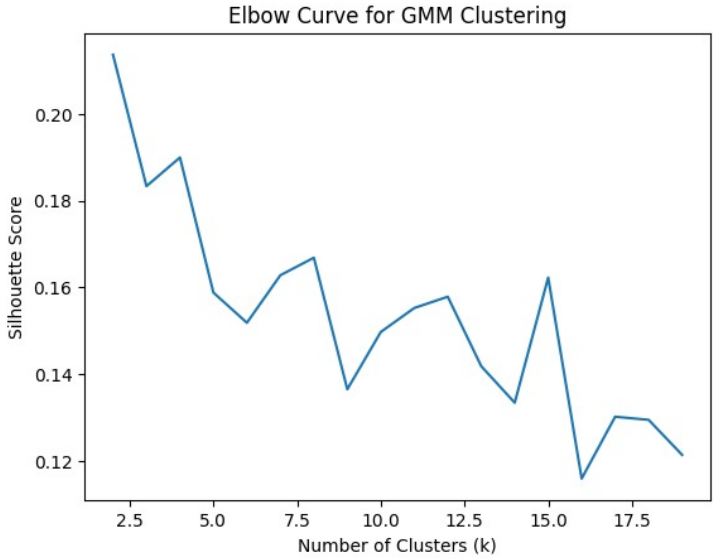

Figure 12: Gaussian Mixture Model

The GMM clustering helps uncover distinct behavioral segments within your customer base. Recognizing these groups enables your team to predict churn more accurately by tailoring specific interventions for each cluster.Interpretation for CLTV Prediction: The GMM provides insight into which customer segments are more valuable over time. By analyzing spending patterns, usage, and service engagement within each cluster, you can model the lifetime value and prioritize segments that contribute more to long-term profitability.

K Means Clustering

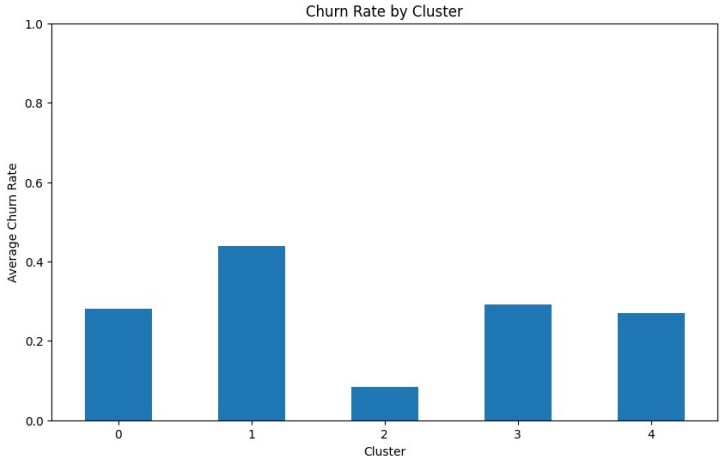

Figure 13: K Means Clustering

K-Means complements the churn prediction by providing a clear segmentation of customers into distinct groups. Identifying attributes that are common in clusters with higher churn rates can inform more targeted churn prevention strategies.Interpretation for CLTV Prediction: By integrating K-Means clustering results, you can segment customers not only by their risk of churning but also by their potential value. It helps in focusing retention efforts on high-value customers and customizing service offerings to enhance their lifetime value.

Survival Analysis

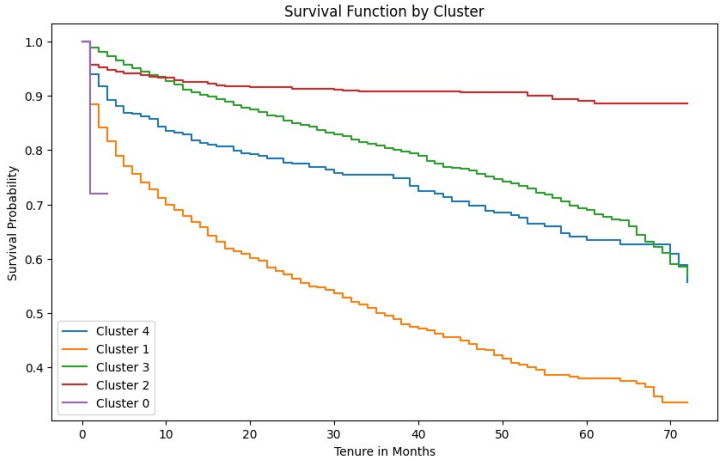

Figure 14: Survival Analysis

The Kaplan-Meier estimator's survival curves offer a dynamic perspective on churn. They show how long customers typically continue with your services before churning, which is crucial for preemptive retention actions.Interpretation for CLTV Prediction: Survival analysis directly impacts CLTV by revealing the expected duration of a customer relationship. Understanding the 'survival' probability over time allows you to adjust the CLTV models accurately and focus on extending the tenure of profitable customers.

The survival analysis plot reveals significant differences in churn risk among various customer clusters. Specifically, Cluster 0 displays a rapid decline in survival probability shortly after engagement begins, indicating these customers may be at high risk of early churn, possibly due to dissatisfaction or being in a trial phase. In contrast, Cluster 2's consistently low survival probability across all tenures marks it as having the highest churn risk, signaling a need for focused retention strategies. In comparison, Clusters 1, 3, and 4 exhibit more gradual decreases in survival probability, suggesting a more robust retention within these groups, particularly in the early tenure stages. Notably, Cluster 4 demonstrates a remarkable stability over time, implying a low churn rate and potentially high customer loyalty or binding contracts. The clear distinction between the clusters' survival probabilities over time underscores the varied experiences and levels of satisfaction across the customer base, providing actionable insights for tailored engagement and retention initiatives.

Integrated Insights for Churn and CLTV prediction

The models employed in the project collectively establish a nuanced approach to predict both the likelihood and timing of customer churn, as well as their lifetime value. The decision tree model illuminates the factors that may lead to churn. Clustering algorithms segment the customers into meaningful groups, revealing who is more likely to churn. Meanwhile, survival analysis provides the temporal dimension, indicating when churn is likely to occur.

The insights gleaned from these models facilitate the formulation of targeted strategies for customer retention and value optimization. Strategies such as loyalty programs may be tailored for high-value customer segments at risk of churn. Similarly, communication strategies can be customized for different clusters to enhance customer engagement, thereby potentially increasing their lifetime value.

In conclusion, the models implemented in this project do more than predict churn; they are instrumental in comprehending and enhancing customer lifetime value. Each model offers distinct insights which, when integrated, contribute to a comprehensive strategy aimed at improving customer retention and maximizing revenue.