PREDICTIVE CUSTOMER ANALYTICS

Customer Lifetime Value (CLV) & Behavioral Segmentation

WIDS 2025

Midterm Project Report

31st January 2026

Name: Onkar Patil
Roll Number: 23B0643
Mentor: Harshul Jain, Sneha Mane

Contents

1 Abstract

The implementation of retention strategies has proven difficult in a high competition, low switching cost environment, exemplified by telecommunications subscription businesses. With the majority of research indicating a greater cost associated with the acquisition of new customers than the retention of existing customers, the ability complete to identify the risk of customer churn becomes important from a business perspective.

In this regard, a customer analytic framework is established that is capable of predicting churn through the use of a variety of processes, including data pre-processing, exploratory data analysis, RFM pivot behavioral segmentation, customer churn machine learning models prediction, and customer lifetime value (LVC) estimation. Customer demographics, service subscriptions, and billing data, in addition to the Telco Customer Churn Dataset, are utilized for the behavioral analysis of customers exhibiting churn. Churn probabilities are predicted using the Logistic Regression and Random Forest models, while CLV analysis provides a means to reorder customers based on value.

The actionable insight provided in this project aims to suggest integrated analytical modeling with business metrics to facilitate focused retention efforts and data informed strategies.

2 Introduction

The last few years have shown a significant increase in the adoption of data driven business strategies which manage the customer relationships. For paid subscription based services such as telecommunication , customer churn represents a large financial challenge. Customers have various altered available, and dissatisfaction with the high pricing, service quality can lead to rapid loss in customers.

The conventional way of customer retention by organizations has been a reactive process, where the organization reacts to the dissatisfaction of the customer only after the customer has left. This process has been inefficient and has resulted in unnecessary losses of revenue. Predictive customer analytics assists organizations in shifting from a reactive process to a proactive process of customer retention by predicting the customers who are likely to leave and taking necessary steps before the termination of services.

The objective of this project is to develop a predictive customer analytics pipeline that integrates statistical analysis, machine learning, and business insights. By integrating the churn prediction model in the behavioral segmentation and CLV estimation , the analysis moves further prediction only and focus on actions which are the insightful and that can guide marketing and the retention strategies.

3 Methodological Framework

The project adopts a systematic analytical framework that is generally applied in industry-level customer analytics studies. Rather than considering the prediction of churn as a machine learning task, the methodological framework of this project combines exploratory analysis, behavioral segmentation, value analysis, and prediction into a single process..

The project begins by working with the raw customer data and cleaning it so that missing or incorrect values do not affect the analysis. This step is important because poor data quality can lead to misleading results later on. After cleaning the data, exploratory analysis is performed to understand how customers behave and how different factors relate to churn. Charts, graphs, and summary statistics are used to observe trends and gain basic insights. Based on what is learned during this stage, new variables are created to better describe customer activity and spending behavior.

Once the data is better understood, customers are grouped using the RFM method, which looks at how recently customers interacted with the service, how often they did so, and how much they spent. This approach helps compare customers based on their actual behavior rather than just demographic information. Customer Lifetime Value is then calculated to estimate the long-term contribution of each customer. After that, churn prediction models are developed to identify customers who are more likely to leave. The results from these models are then used to suggest practical actions that businesses can take to reduce churn.

4 Problem Definition and Objectives

This project focuses on the problem of early-stage identification of customer churn and determining prioritization of customers for retention. Understanding customers' value and predicting churn is essential in order to allocate and optimize value in prediction.

This project aims to answer the questions regarding customer behavior through the analysis of the available demographic, service usage, and billing information.

Additionally, the intricate patterns, if any, related to churn are explored in the datasets. Using the RFM framework, customer segmentation is done to ascertain varying levels of engagement. Furthermore, the estimation of customer lifetime value is done to ascertain the potential value of revenue over time. Lastly, churn prediction models such as Logistic Regression (LR) and Random Forest (RF) are constructed and the insights gained are analyzed for practical recommendations to the business.

5 Dataset Description

Telco Customer Churn dataset is utilized in this project. It is public and available for studying customer churn, as it contains customer related data, including service usage and churn status. It has 7043 records – one for each customer.

The dataset has a wide array of information which can be split into three categories. The first one is demographics, which includes gender, whether the customer is a senior citizen, and information related to a customer's partner and dependents. The second one is service used by the customer, i.e., type of phone service, internet service, and any additional services, such as online security and tech support. The last one is billing and account information, which is still related to services used, and includes the customer's duration of service usage (tenure), type of contract, payment method, monthly and total charges.

Since the dataset has a well defined churn column, it makes the application of supervised learning techniques easier on this dataset. Additionally, it makes this dataset appropriate for business case usage, to analyze customer behavior and to understand the customers who are most at risk for attrition.

6 Data Cleaning and Preprocessing

Whenever we handle real world customer datasets, we need to make sure we handle the missing data, incorrect formatting, and any inconsistencies data we might encounter. Because of this, data cleanig and preprocessing has to be done before any relevant analysis and modeling can be done.

The most notable example of incorrect data in the dataset was in the TotalCharges column. Because of formatting issues, this column was treated as text. It was later converted into numeric format. Then, the rows with missing or incorrect numeric values were removed. This was done to prevent analysis errors. Additionally, customer ID columns were also removed as those columns were trivial to the analysis. In order to use machine learning, categorical columns were converted into numeric columns via one hot encoding. The columns with numeric values such as tenure and charges were adjusted so their values are proportionate to one another. This creates better results by minimizing the bias introduced by differences in the values.

The basic steps of preprocessing carried out in this project are available in the code snippet below.

```
df[’TotalCharges’] = pd.to_numeric(df[’TotalCharges’],
errors=’coerce’)
df.dropna(inplace=True)
df[’Churn’] = df[’Churn’].map({’Yes’:  1, ’No’:  0})
df.drop(columns=[’customerID’], inplace=True)
```

Theses steps gives a clean and well formatted dataset which suitable for Exploratory data analysis and the predictive learning

7 Rationale Behind Preprocessing Decisions

The preprocessing steps were done to make the data usable and easy to understand. The TotalCharges column was changed into numeric form because it could not be used properly in its original format. Without this change, doing any money related analysis would not be possible. Rows with missing values were removed instead of filling them because the number of such rows was small and filling them could change the data in a wrong way.

The customer ID column was removed because it is only used to identify customers and does not help in predicting churn. Keeping this column could confuse the model. Categorical values were converted into numbers so that the models can work with them and understand differences in services.

Numerical values were scaled so that no single column has very large values compared to others. This helped the model learn better, especially in Logistic Regression. These steps helped improve model performance and made the results easier to read.

8 Exploratory Data Analysis

EDA was done to look at the data and understand customer behavior. This step helped in finding patterns related to churn and also helped in deciding what to use for modeling. At the start, it was seen that the churn data was not balanced. Around 26.5 % of customers had churned. Because of this, accuracy alone was not enough to judge the model. Other measures like precision, recall, and F1 score were also used.

When contract types were checked, customers with month to month contracts showed much higher churn. Customers with long contracts were more likely to stay. This shows that contract length affects churn. When looking at tenure, most churn happened in the first 12 months. After that, customers usually stayed longer.

Money related factors also affected churn. Customers paying higher monthly charges had higher churn. This was more common for customers using fiber optic internet, which may mean they were not happy with the service compared to the cost.

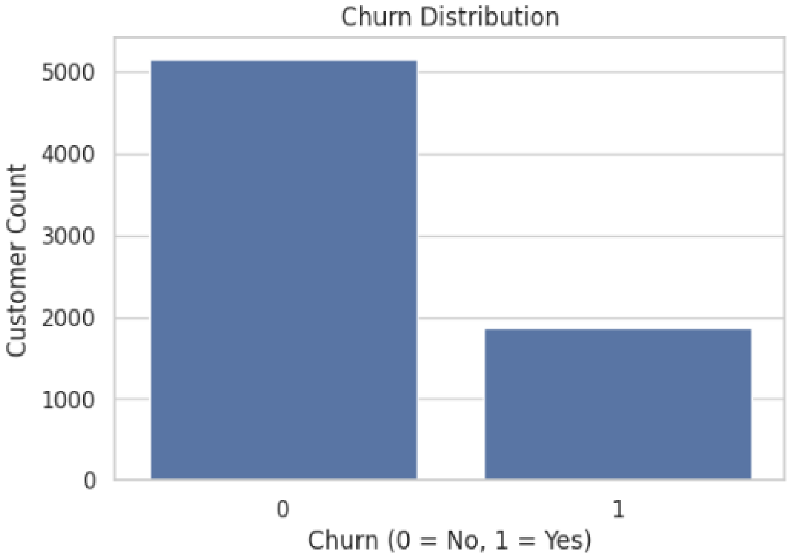


FIG 1.Churn Distribution

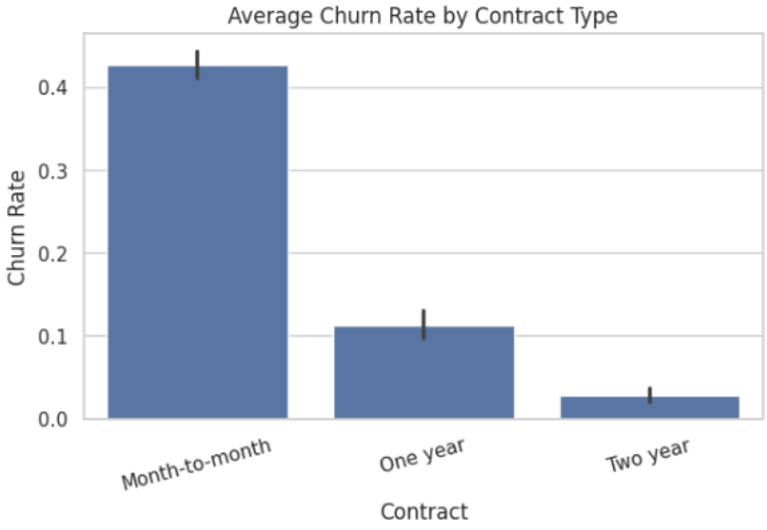


Figure 2: Churn Rate by Contract Type

9 Behavioral Interpretation of Exploratory Findings

The exploratory analysis helped in understanding customer behavior, not just numbers. Customers with month to month contracts had higher churn, which shows that even though flexible plans look good at first, customers may find it easier to leave since there is no strong commitment. Customers with long contracts usually stay longer because they get used to the service and feel more connected to it.

The link between tenure and churn shows that the first few months are very important. Customers who stay after the early period usually continue using the service. Over time, they get comfortable with it and small issues do not affect them as much.

Customers with higher monthly charges showed more churn. This may mean they are more sensitive to price or they feel the service is not worth the cost. Overall, churn does not happen because of one reason only. It happens due to a mix of contract type, customer behavior, and cost related factors.

10 Feature Engineering

Feature engineering was done to make the data more useful for prediction. New features were created from existing data to better represent customer behavior. Tenure related features were used to show how long a customer has stayed with the service. Billing related features were used to show how much value a customer brings.

Service related features were encoded so the model could learn from customer usage patterns. These features together helped give a clearer picture of customer behavior and were useful for churn prediction and segmentation.

11 Why RFM Segmentation is Effective in Practice

RFM segmentation works well because it matches how businesses usually think about customers. Instead of looking only at age or gender, it looks at what customers actually do. This makes the results more useful and easier to act on.

Recency shows how recently a customer used the service, which can help in spotting customers who may leave soon. Frequency shows how often a customer uses the service, and Monetary value shows how much money they spend. When these three are combined, it gives a clear overall picture of the customer.

In this project, RFM segmentation helped connect basic analysis with prediction. It made it easier to decide which customers need retention offers and which customers could be targeted for additional services.

12 RFM-Based Customer Segmentation

Customer segmentation was done using the RFM method. This method is commonly used because it is easy to understand and useful for business decisions.

Recency measures how recently a customer interacted with the service. Frequency shows how often the customer uses it. Monetary value shows how much money the customer contributes. Unlike demographic methods, RFM directly focuses on customer actions.

Customers were given scores for each RFM part using ranking based on the data. These scores were combined to get an overall RFM score. Based on this score, customers were grouped into segments such as high value customers, customers at risk of leaving, and customers who may become loyal.

```
rfm[’R_Score’] = pd.qcut(rfm[’Recency’], 4, labels=[4,3,2,1])
rfm[’F_Score’] = pd.qcut(rfm[’Frequency’], 4, labels=[1,2,3,4])
rfm[’M_Score’] = pd.qcut(rfm[’Monetary’], 4, labels=[1,2,3,4])
```

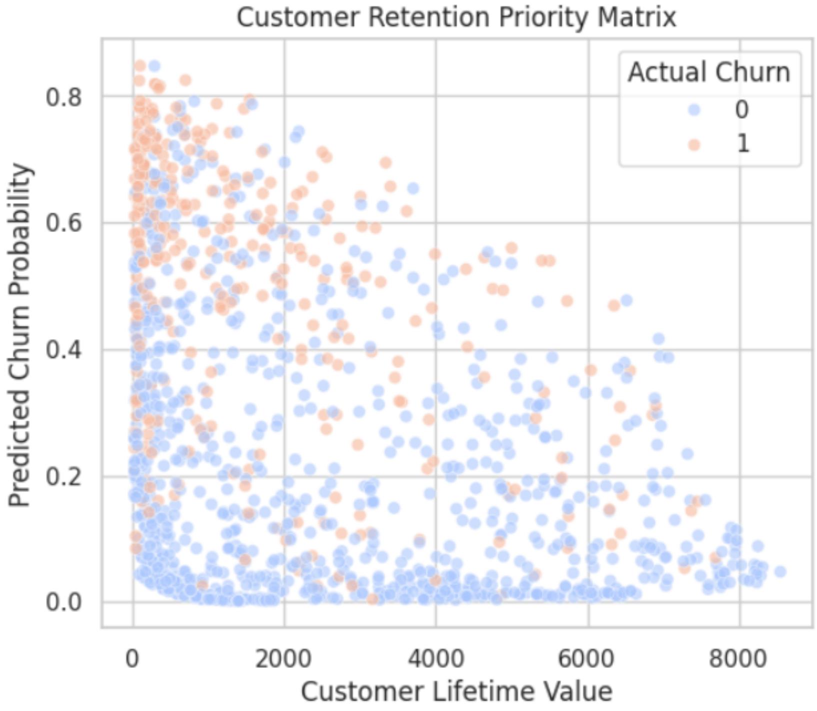
These code snippets provide a base for targeted retention and marketing strategies.

13 Customer Lifetime Value Analysis

Customer Lifetime Value or CLV shows how much money a customer is likely to bring in during the time they stay with the company. Churn prediction tells us who might leave, but CLV tells us which customers are more important to keep.

In this project, CLV was calculated in a simple way using tenure and monthly charges. This is not perfect but it gives an idea of how valuable each customer is over time.

The analysis showed that customers with high RFM scores usually have higher CLV. This means that the RFM segmentation works well and helps to focus on keeping the most valuable customers.



14 Churn Prediction Models

Two models were used to predict customer churn. Logistic Regression was used as a basic model because it is easy to understand and works well for yes/no problems. Random Forest was used as a second model because it can handle more complex patterns and interactions between features.

The data was split into training and testing sets to see how well the models work on new data. Both models were trained using features from customer details, service usage, billing, RFM scores, and CLV.

15 Interpretability and Model Trust

In real business, it is important to understand why the model says a customer might leave. Logistic Regression was used because it is simple and we can see how each feature affects the prediction. This helps business people trust the model.

Random Forest is more flexible and can capture complicated patterns, but it is harder to understand. That is why it was used together with Logistic Regression, not instead of it. Using both models helps balance understanding and performance so that the results can be used in real decisions.

16 Model Evaluation and Performance

The model is checked by the parameter like Accuracy, precision ,recall, F1- score. Logistic regression gave good results and was easy to explain. Random Forest perform better on precision. Implying it can find the more customers likely to leave.

Recall was important in this project to avoid missing customers who might churn. The goal was to make sure risky customers are not overlooked.

```
results_df = pd.DataFrame({})
"Model":  ["Logistic Regression", "Random Forest"],
"Accuracy":  [0.80, 0.79],
"Precision":  [0.61, 0.62],
"Recall":  [0.55, 0.56],
"F1-Score":  [0.58, 0.59]
}
```

model	accuracy	precision	recall	F1-score
Logistic regression	0.80241648898	0.6463414634	0.566844919	0.603988603
Random Forest	0.791755507	0.637288133	0.5026737967	0.5620328849

17 Error Analysis and Business Implications

The model errors were checked from a business point of view. False negatives, which are customers predicted to stay but actually leave, are worse because they mean lost chances to keep customers. That's why Recall was considered more important.

False positives, customers predicted to leave but actually stay, are not as bad. They can be handled with cheap retention actions. This is why Recall and F1-score matter more than just Accuracy. Looking at the errors like this helps make sure the model results match business goals and acceptable risks.

18 Business Insights and Recommendations

Using churn predictions, RFM segments, and CLV together helps in deciding what to do with customers. Customers who have high CLV and high risk of leaving should be contacted first with offers or upgrades. Customers with low CLV and high risk can be handled with cheaper methods. High value customers who are low risk can be targeted for extra services or loyalty programs. This shows how analytics can be used in actual business decisions and not just in theory.

19 Limitations

There are some limitations in this analysis. The dataset is just a snapshot and does not show real-time behavior. The CLV calculation is simple and does not use probability models. Also, things like customer feedback or sentiment were not included. There are some limitations in this analysis. The dataset is just a snapshot and does not show real-time behavior. The CLV calculation is simple and does not use probability models. Also, things like customer feedback or sentiment were not included.

20 Future Scope

This project gives a basic and working setup for predictive customer analytics, but there is still a lot that can be improved in the future. One major improvement can be using more advanced machine learning models. Models like XGBoost, Gradient Boosting, or survival analysis can be tried to get better churn prediction results. These models can handle complex patterns and time based churn better than simple models.

CLV calculation can also be improved. In this project, CLV was calculated in a simple way, but in future probabilistic models like BG/NBD and Gamma-Gamma can be used. These models consider uncertainty in customer behavior and buying patterns, which can give a better idea of long term customer value instead of only using past data.

Another area for improvement is real time analysis. Dashboards can be created that update churn risk and CLV scores regularly or almost in real time. This will help marketing and customer teams react faster when customers show signs of leaving. It also helps teams stay on the same page.

Finally, the full system can be connected to CRM platforms. Once connected, actions like offers, follow ups, or support calls can be triggered automatically based on model results. This will make churn prediction part of daily business work instead of a one time analysis.

Overall, these future improvements can make the system more accurate, more useful, and easier to use at a larger scale.

21 Deployment Considerations

In real business use, the analytics system made in this project can be added into existing CRM systems. This helps the company use the results in daily work, not just in reports or files.

Churn chances and CLV values can be updated again and again, like every week or every month. New customer data is used for this. These values can be shown in dashboards so teams can see them easily. Marketing and customer teams can quickly find customers who are risky and important.

Automation helps a lot here. Instead of checking data manually, the system can watch customer behavior on its own. If a customer starts showing churn signs, it can be flagged early. This helps the business act fast, like giving offers or fixing service issues.

This also helps teams work together better. Marketing can plan campaigns using the data. Support teams can focus on customers who need help urgently. Managers can see overall numbers and make better decisions.

Overall, putting analytics into CRM systems makes churn prediction actually useful. It becomes part of normal business work. This helps companies react faster, treat customers better, and grow steadily over time.

22 Ethical and Practical Considerations

Predictive customer analytics can help businesses a lot, but it must be used carefully. How customer data is used affects how customers are treated and targeted, so it is important to use it in the right way.

One big issue is customer privacy. Companies use a lot of data like personal details, behavior, and payment information to build churn models. While this helps prediction, collecting too much data or using it wrongly can break customer trust. Businesses must follow data protection rules and clearly explain how customer data is collected and used.

Another problem is bias in models. Models learn from past data, and this data may already contain unfair treatment in pricing or service. If this is not checked, the model may treat some customer groups unfairly. Because of this, models should be checked regularly to make sure they are fair for all customers.

From a practical side, too much retention effort can be harmful. If customers are contacted too often or given too many offers, they may get annoyed instead of staying. Churn predictions should be used carefully so customers are not disturbed too much.

Model performance can also change over time. Customer behavior changes because of competition, market changes, or new services. If models are not updated, they may stop working well. So models should be retrained and checked from time to time.

Finally, predictive models should support human decisions, not replace them. Models give useful information, but human judgment is still needed. Using both together helps businesses make better and more responsible decisions.

23 Conclusion

This project shows a full end to end approach for predictive customer analytics with a focus on customer churn. The work started with cleaning and preparing the data so that it could be used properly. After that, exploratory analysis was done to understand customer behavior and find factors that affect churn. RFM segmentation was used to study how customers engage with the service, and CLV was calculated to understand which customers are more important in the long run.

Based on these results, machine learning models were built to predict churn. Logistic Regression was used as a simple and easy to understand model, while Random Forest was used to capture more complex patterns. The results showed that using behavior and financial features together gives better churn prediction than using them separately.

The project did not focus only on prediction results but also on business use. By combining churn scores with CLV and segmentation, it became easier to decide which customers should be targeted first for retention. This helped connect data analysis with real business actions like marketing and customer support.

Overall, this project shows that predictive analytics is not only about building models. When used correctly, it can help companies keep customers, use resources better, and build stronger customer relationships over time.

24 References

1. Telco Customer Churn Dataset – Kaggle
2. Statology: Customer Lifetime Value with Python
3. CleverTap: RFM Analysis
4. BCIIT e-Journal (2025): Customer Churn Prediction using Machine Learning