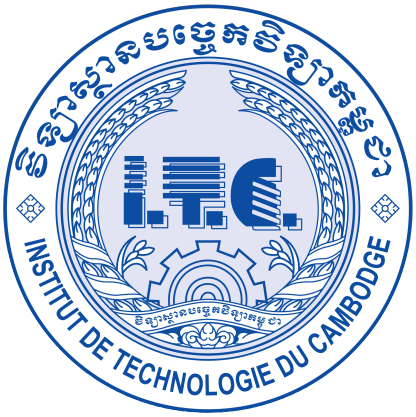

Institute of Technology of Cambodia

Regression Analysis
2022-2023

3rd-year Engineer's Degree in Data Science

Department of Applied Mathematics and Statistics

Predicting customer churn using logistic regression

Group members:

Name	ID
Ya Manon	e20200745
Ngeav Bonat	e20201691
Chhon Chaina	e20200934
Set Sophy	e20201576
Hun Sokrarith	e20201218
Phal Davy	e20201437

Lecturers:
Dr. Phauk Sokkhey
Mr. Nhim Malai

Submission Date: 18.June.2023

Contents

Abstract

This study uses logistic regression to predict customer churn in a telecommunications company. The dataset includes information on customer demographics, usage patterns, and account information. Logistic regression is used to model the relationship between these variables and whether or not customers churned. The study finds that several variables are significant predictors of churn, including contract type, monthly charges, and tenure. The logistic regression model has an accuracy of 79% indicating that it can be useful in identifying customers who are at risk of churning. Moreover, L1 regularization is also used in the project to prevent overfitting and improve the generalization performance of the model, which has an accuracy of 80%İn conclusion, we have some recommendations, businesses can enhance customer satisfaction, loyalty, and untimately reduce churn rates.
Key Words: churn; logistic regression; L1 regularization; predict; accuracy

1 Introduction

In simple words, customer attrition occurs whenever a customer stops doing business with a service provider or a company. Churners have always been a big issue for any service-providing company. Churning increases the cost of the company as well as decreases its rate of profit. Considering the machine learning perspective, the churn prediction is supervised and that can be defined as Given a predefined forecast horizon, the goal is to predict the future churners over that horizon, given the data associated with each subscriber in the network. Churn prediction aims to identify subscribers who are about to transfer their business to a competitor. Since the cost associated with customer acquisition is much greater than the cost of customer retention, churn prediction has emerged as a crucial Business Intelligence (BI) application for modern service-providing companies. The flow of raw materials or other goods to end customers is called as logistics. The essential success factor for any logistics industry lies in delivering items to the correct place and at the appropriate time with a reasonable cost. Customer dissatisfaction at any of the stages of this process leads to a huge loss in the business and that is where the exact concern lies. A long-term relationship with the customers is a very crucial factor in the logistics industry because of the innumerable aspects of service encounters that can easily be imitated by competitors. One of the gauging successes in the logistics industry is customer churn. Therefore, there is a huge need for a defensive marketing strategy that prevents customers from switching service providers. Customer churn causes revenue loss and other negative effects on corporate operations. Therefore, our idea mainly focuses on the customer churn prediction model for identifying the key factors which are crucial and which cause the churn. The set of techniques that we use to do the same includes Logistic regression(LGR), decision tree analysis, and artificial neural network.

1.1 Goal

The primary goal of this project is focused on helping business owners find services in the company which is good and bad for our customers we use a machine learning model and choose the Logistic Algorithm. The second goal is to consider the possibility of introducing self-learning models and improving their accuracy based on new data. To reach the goals of the project, it is required to address the following questions:

- What does churn stand for?
- Why do businesses want to prevent churn?

1.2 Setup

To analyze the Telecommunication customer churn dataset and develop a machine learning model using the Logistic Algorithm, the following setup is required:

1. Obtain the Telecommunication customer churn dataset: The dataset contains valuable information about a fictional telco company operating in California during Q3. It includes details on demographics, service usage, satisfaction scores, and customer lifetime value.
2. Data preprocessing: Before applying the machine learning model, it is necessary to preprocess the dataset. This involves handling missing values, encoding categorical variables, scaling numerical features, and splitting the data into training and testing sets.
3. Feature selection: To identify the factors contributing to customer churn, it is important to perform feature selection. This can involve techniques such as correlation analysis, feature importance ranking, or domain knowledge to select the most relevant features for the model.

4. Model development: Implement the Logistic Algorithm, a popular classification algorithm suitable for predicting binary outcomes like churn. Train the model using the preprocessed data and evaluate its performance using appropriate metrics such as accuracy, precision, recall, and F1-score.

5. Interpretation and analysis: Once the model is developed and evaluated, analyze the results to uncover patterns, correlations, and predictive indicators that contribute to customer churn. This analysis will help in understanding the impact of various factors and identifying strategies for improved retention.

6. Consideration of self-learning models: To improve the accuracy of the model over time, consider the possibility of introducing self-learning models. These models can continuously learn from new data and adapt their predictions, allowing for better churn prevention and customer satisfaction.

2 Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to analyzing and understanding data sets. EDA involves generating summary statistics for each variable in the data set, as well as creating visualizations such as histograms, scatterplots, and heatmaps to identify patterns, trends, and relationships among variables. The goal of EDA is to gain insights into the data, understand its underlying structure and characteristics, and identify potential issues or anomalies that may need further investigation. EDA is often conducted at the beginning of a data analysis project before more complex modeling or statistical techniques are applied to the data.

2.1 Data Description

The Telco customer churn data contains information about a fictional telco company that provided home phone and internet services to 7043 customers in California in Q3. It indicates which customers have left, stayed, or signed up for service. Multiple important demographics are included for each customer, as well as a Satisfaction Score, Churn Score, and Customer Lifetime Value(CLTV) index.

- **customerID:** a unique ID that identifies each customer
- **gender:** whether the customer is a male and female
- **SeniorCitizen:** indicates if the customers are 65 or elder (1, 0)
- **Partner:** indicates if the customer has a partner or not (Yes, No)
- **Dependents:** whether the customer lives with any dependents (children, parents, grandparents) (Yes, No)
- **tenure:** number of months the customer has stayed with the company
- **PhoneService:** indicates if the customer subscribes to home phone service with the company (Yes, No)
- **MultipleLines:** whether the customer has multiple lines OR NOT (Yes, No)
- InternetService: customer's internet service provider (DSL, Fiber Optic, No)
- **OnlineSecurity:** whether the customer has online security provided by the company (Yes, No, No internet)
- **OnlineBackup:** whether the customer has online backup or not (Yes, No, No internet service)
- **DeviceProtection:** whether the customer has a device protection plan for their internet equipment (Yes, No, No internet service)
- **TechSupport:** whether the customer has tech support or not (Yes, No, No internet service)
- **StreamingTV:** whether the customer has streaming TV or not (Yes, No, No internet service)
- **StreamMovies:** whether the customer used their Internet service to stream movies or not (Yes, No, No internet service)

- **Contract:** indicates the customer's current contract type (Month-to-month, One year and Two year)
- **PaperlessBilling:** whether the customer has paperless billing or not (Yes, No)
- **PaymentMethod:** indicates the payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)
- **MonthlyCharges:** the amount charged to the customer monthly
- **TotalCharges:** indicates the customer's total charges
- **Churn:** indicates whether the customer churn (Yes, No)

2.2 Data Cleaning

2.2.1 Missing value

The real world doesn't have a perfect dataset. As we know missing and it's important to address this issue in machine learning projects. Missing data can occur due to various reasons such as human error in data entry, incomplete surveys, or sensor failure. One approach to dealing with missing data is to remove the observations with missing values, but this can lead to the loss of valuable information and reduced sample size. To avoid any problem with missing data, we need to perform any technique but, in our project use mean, mode and median for this project.

2.2.2 Duplicated

Duplicated data refers to instances where identical or very similar data appears in multiple locations within a database or data set. In this case, we could not find any duplicates out of 7023 rows.

2.2.3 Outlier

In statistics, an outlier is a data point that differs significantly from other observations in a dataset. Outliers can be caused by measurement or recording errors, or they may indicate a rare event or a significant deviation from the normal pattern of the data. Outliers can have a significant impact on statistical analysis and can lead to incorrect conclusions if not carefully accounted for. Identifying and handling outliers is an important step in many data analysis tasks, including regression analysis, hypothesis testing, and clustering analysis. In this case, we had check the outlier for 4 features including SeniorCitizen, tenure, MonthlyCharges and TotalCharges as you can see in figure 1.

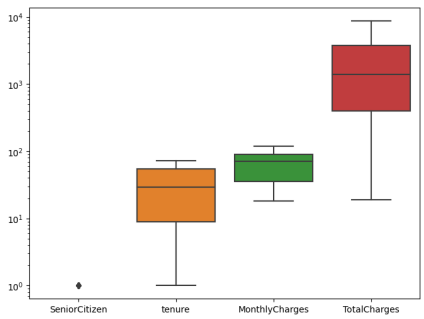

Figure 1: The dataset without the outlier.

2.3 Data Visualization

Data visualization is the graphical representation of data and information. It involves creating visual representations of complex data sets in order to help people understand and interpret the data more easily. This can include charts, graphs, maps, diagrams, and other forms of visualizations that display quantitative or qualitative data. The goal of data visualization is to communicate information in a way that is clear, concise, and meaningful to the intended audience. By presenting data visually, it allows for patterns, trends, and relationships to be quickly identified and understood, which can lead to better decision-making and insights. First, we plot the bar plot to see which feature is very effective on churning. First, we plotted for the customer and we can see in figure 2:

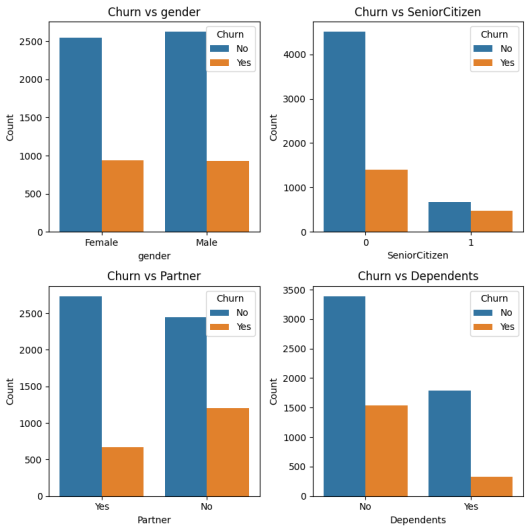

Figure 2: Churn versus Customers

- Gender has no influence on whether the customer will churn or not.
- Non-senior citizen churn more than non-senior in absolute term but in relative term senior citizens churn more often.
- Customers without partner churn more often than their counterparts.
- Customer without dependants churn more than customers with dependents.

Second, we plot all the services versus churn in case to check whether which one is the most churning as you can see in the following figure 3:

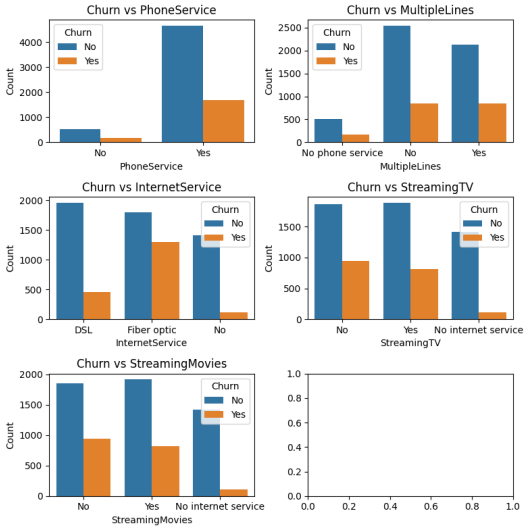

Figure 3: Churn versus Services

- Streaming TV, streaming Movies, MultipleLines have no impact on churn rate because we can see that churn is less than 50% of not churn. So, we can make inference that all of those have not empact on churning.
- Customer who have fiber optic tend to churn significantly more than ones that have DSL has lower chrun rate.

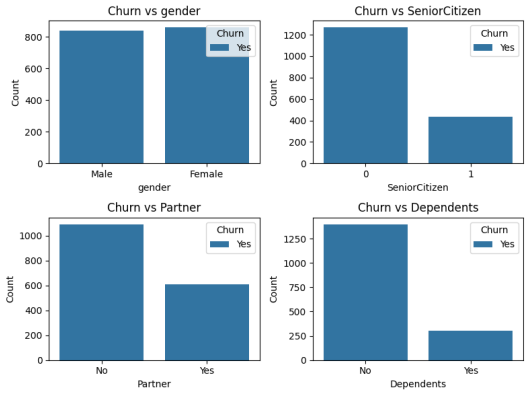

Figure 4: Count customer versus churn

We can conclude that customer who have churned and use phone service are likely to be non-seniors, without partner and dependents. We had got the same conclusion when analyzing the same variable regardless of customer's services.

Third, we plot the security support including online security, online backup, device protection and tech support versus churn in the figure 5:

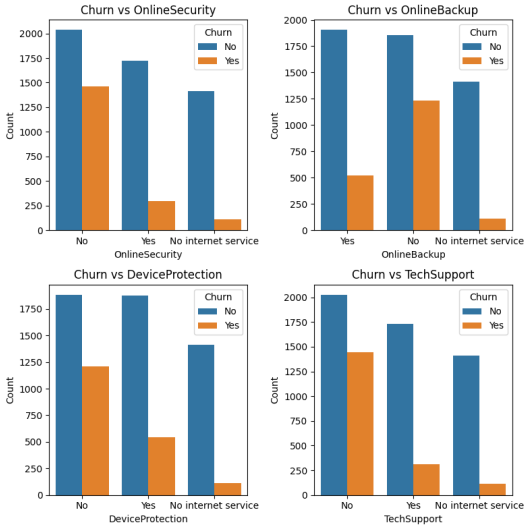

Figure 5: Churn versus Security support

Fourth, we plotted the contract type versus churn in figure 6:

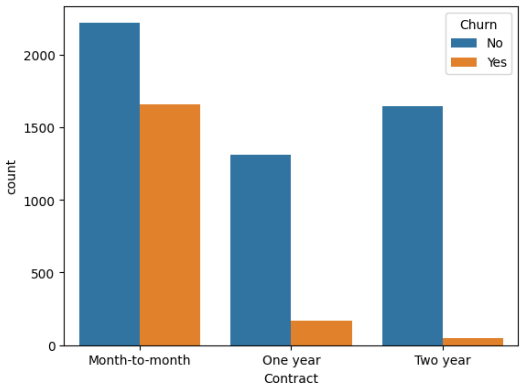

Figure 6: Churn versus Payment method

Customers with Month-to-month contracts tend to churn significantly more often than customers with one year and two year contracts.

Fifth, we plotted the payment option and options versus churn in figure 8:

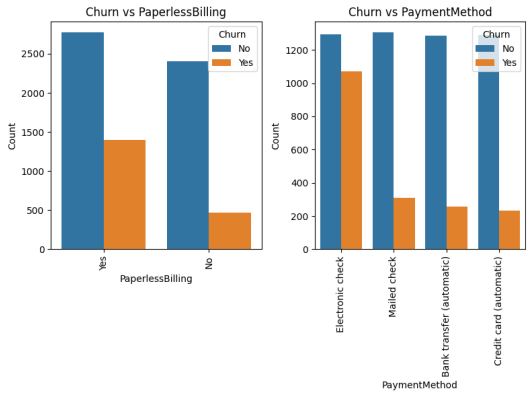

Figure 7: Churn versus Payment Method

Lastly, we plotted the tenure group versus churn in figure 9:

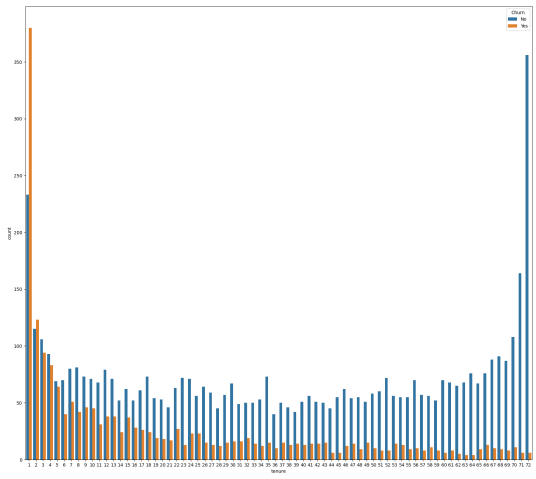

Figure 8: Churn analysis based on tenure group

Customers who spent less than 1 year using our services tend to churn substantially more than the customer of the tenure groups. As the tenure increases, the churn rate decrease.

3 Feature Engineering

Feature engineering is the process of selecting, transforming, and creating features from raw data to improve the performance of machine learning models. It involves extracting relevant information from the data and representing it in a format that can be effectively utilized by the model.
In this case, we divided the data into 3 groups such as numeric features, yes/no columns, and categorical features.

3.1 Data Prepocessing

3.1.1 Scaling

We used **Min-Max scaling**, also known as normalization, is a common technique used to rescale numerical features in a dataset. It transforms the values of the features to a fixed range, typically between 0 and 1, based on the minimum and maximum values in the original data.

3.1.2 Handle Feature Categorical

1. **Label Encoding:** Label encoding is a technique used to convert categorical variables into numerical representations. In machine learning, many algorithms require numerical input, so label encoding is often employed to transform categorical data into a format that can be processed by these algorithms.

   In label encoding, each unique category or label in a categorical feature is assigned a unique numerical value. The process involves the following steps:

   (a) Identify the categorical feature that needs to be encoded.

   (b) Assign a numerical value to each unique category in the feature. The values can be assigned arbitrarily or based on some logical order.

   (c) Replace the original categorical values with their corresponding numerical values.

   we used this method for convert yes into 1 and no into 0

2. **One-Hot Encoding:** is a technique used to represent categorical variables as binary vectors. It is commonly employed when dealing with categorical data in machine learning, as many algorithms require numerical input.

   In One-Hot Encoding, each unique category or label in a categorical feature is transformed into a separate binary feature or variable. For a feature with 'n' unique categories, 'n' binary variables are created, where each variable represents one category. The process involves the following steps:

   (a) Identify the categorical feature that needs to be encoded.

   (b) Create 'n' binary variables, one for each unique category in the feature.

   (c) Assign a value of 1 to the corresponding binary variable if the category is present for a particular observation, and 0 otherwise.

   For example in this project, consider a categorical feature "Partner" with the following values: "Yes", "No". After one-hot encoding, the feature may be represented as three separate binary variables: "1", and "0".

   n this case, we converted categorical variables into numerical format

3.2 Feature selection

Feature selection is a crucial step in the machine learning pipeline that involves identifying and selecting the most relevant features from a given set of input features. The objective of feature selection is to improve model performance, reduce overfitting, and enhance interpretability by focusing on the most important features that contribute to the prediction task.

Let $\mathbf{X}$ denote the matrix of input features with dimensions $n \times m$, where n is the number of samples and m is the number of features. The target variable is denoted as $\mathbf{y}$.

There are various techniques available for feature selection, including filter methods, wrapper methods, and embedded methods. In this study, we employed Recursive Feature Elimination (RFE) in conjunction with logistic regression.

RFE is an iterative feature selection method that eliminates less important features based on their contribution to the model's performance. It considers feature interactions during the elimination process, allowing for the capture of dependencies and interactions among features.

To perform feature selection, we applied RFE with logistic regression as the estimator. Let $\mathbf{w}$ represent the weight vector and b be the bias term of the logistic regression model. The RFE process involves iteratively eliminating less important features by minimizing the logistic regression loss function:

$$\min_{\mathbf{w},b} \frac{1}{n} \sum_{i=1}^{n} \left(\log(1 + \exp(-\mathbf{y}_i(\mathbf{w}^T\mathbf{x}_i + b)))\right) \tag{1}$$

where $\mathbf{x}_i$ represents the feature vector for sample i and $\mathbf{y}_i$ is the corresponding target value.

The elimination process continues until a specified number of features remain or a stopping criterion is met. The importance of each feature is evaluated based on its contribution to the model's performance.

The advantage of using logistic regression with RFE is its ability to select the most informative features while considering feature interactions. By focusing on relevant features, we aim to improve model efficiency, reduce overfitting, and enhance the interpretability of the results.

The selected subset of features obtained from RFE with logistic regression will be used as input for model training and evaluation in subsequent sections of this study.

4 Method

4.1 Preliminary Analysis

In the initial model summary generated by stats models. stats. outliers influence the import variance inflation factor model, we see in the Notes that there could be strong multicollinearity in the model. We know from the correlation charts that Tenure is highly correlated. It is then necessary that use either tenure in Logistic Regression analysis.

4.2 Logistic Regression

The dataset should contain features (columns) that can be used to predict customer churn, along with a target variable column named "Churn" indicating whether a customer has churned (1) or not (0).
The dataset is then split into features(X) and the target variable(y) using the drop method. The data is further split into training and testing sets using 'train test split'.

To address our research objective and predict the outcome variable, we employed a logistic regression model. Logistic regression is a popular statistical technique used to model the relationship between a binary dependent variable and one or more independent variables.

The logistic regression model can be represented mathematically as follows:

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

where:

- $\text{logit}(p)$ is the natural logarithm of the odds of the dependent variable (p) being in the positive class.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients or weights assigned to the independent variables $X_1, X_2, \ldots, X_n$.
- $X_1, X_2, \ldots, X_n$ are the independent variables used to predict the dependent variable.

The logistic regression model applies the logistic function, also known as the sigmoid function, to the linear combination of the independent variables. This transformation maps the linear combination to a probability value between 0 and 1.

The logistic function is defined as:

$$p = \frac{1}{1 + e^{-\text{logit}(p)}}$$

where:

- p is the predicted probability of the dependent variable being in the positive class.
- e is the base of the natural logarithm.

The logistic regression model estimates the coefficients ($\beta_0, \beta_1, \beta_2, \ldots, \beta_n$) using maximum likelihood estimation or other optimization techniques. These coefficients provide information about the magnitude and direction of the relationship between the independent variables and the log-odds of the dependent variable.

To interpret the logistic regression model, we examine the coefficients (β) and their corresponding p-values or confidence intervals. A positive coefficient indicates that an increase in the corresponding independent variable leads to increased log-odds of the dependent variable being in a positive class. Conversely, a negative coefficient implies a decrease in the log odds.

Additionally, the odds ratio can be calculated by exponentiating the coefficient estimates. The odds ratio represents the change in odds of the dependent variable for a one-unit increase in the corresponding independent variable. It provides insights into the relative impact of the independent variables on the likelihood of the event occurring.

In our logistic regression model, we performed feature selection and employed techniques such as stepwise regression or regularization methods to identify the most relevant independent variables. We assessed the goodness-of-fit of the model using appropriate evaluation metrics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or deviance.

The assumptions of logistic regression, such as linearity, independence of errors, absence of multicollinearity, and lack of influential outliers, were checked to ensure the validity of the model. We also performed diagnostic tests, such as assessing residuals and influential observations, to evaluate the model's performance and identify any potential issues.

The logistic regression model was implemented using Python (or any other preferred programming language) and the scikit-learn or statsmodels library. The dataset was split into training and testing sets to evaluate the model's performance on unseen data. Cross-validation techniques, such as k-fold cross-validation, may have been employed to further assess the model's generalizability.

Overall, the logistic regression model served as a valuable tool in predicting the binary outcome variable and understanding the relationships between the independent variables and the log-odds of the dependent variable.

4.2.1 Logistic Regression with L1 Regularization

To address our research objective of predicting customer churn, we employed logistic regression with L1 regularization, also known as Lasso regularization. Logistic regression is a widely used statistical technique for modeling the relationship between a binary dependent variable and independent variables.

In logistic regression, the goal is to estimate the probabilities of the binary outcome variable, in this case, whether a customer churns or not. The L1 regularization term, also known as Lasso regularization, is added to the logistic regression objective function to encourage sparsity and feature selection. Lasso regularization imposes a penalty that shrinks the coefficients of irrelevant or less important features towards zero, effectively selecting the most informative features for predicting churn.

The logistic regression model with L1 regularization can be represented mathematically as follows:

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

where:

- $\text{logit}(p)$ is the natural logarithm of the odds of the dependent variable (p) being in the positive class.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients assigned to the independent variables $X_1, X_2, \ldots, X_n$.
- $X_1, X_2, \ldots, X_n$ are the independent variables used to predict customer churn.

The L1 regularization term is added to the logistic regression objective function to promote sparsity. The objective function becomes:

$$\text{minimize } -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log(p^{(i)}) + (1-y^{(i)})\log(1-p^{(i)}) + \lambda\sum_{j=1}^{n}|\beta_j|$$

where:

- m is the number of training examples.
- $y^{(i)}$ is the target variable for the i-th training example.
- $p^{(i)}$ is the predicted probability of the i-th training example being in the positive class.
- λ is the regularization parameter controlling the strength of the L1 regularization.

1. The L1 regularization term, $\lambda\sum_{j=1}^{n}|\beta_j|$, penalizes the absolute values of the coefficients, encouraging some coefficients to become exactly zero. This results in a sparse model, where only the most important features remain, promoting feature selection.
2. To find the optimal values of the coefficients (β), an optimization algorithm such as coordinate descent or proximal gradient descent can be employed. These algorithms minimize the logistic regression objective function with the added L1 regularization term.
3. Implementations of logistic regression with L1 regularization are available in machine learning libraries such as scikit-learn in Python or glmnet in R. These libraries provide efficient and optimized implementations of logistic regression with L1 regularization, making it easy to apply and experiment with different values of the regularization parameter λ.
4. To determine the optimal value of λ, model selection techniques such as cross-validation or grid search can be employed. These techniques involve evaluating the performance of the logistic regression model with different values of λ on validation data, selecting the value that provides the best trade-off between model complexity and generalization performance.

By applying logistic regression with L1 regularization, we can effectively select the most important features for predicting customer churn, improving the interpretability and efficiency of the model.

5 Conclusion and Recommendation

5.1 Results and Recommentdation

The coefficients represent the estimated effect of each feature on the probability of customer churn. Positive coefficients indicate that an increase in the corresponding feature value increases the probability of churn, while negative coefficients indicate a decrease in the probability of churn. The magnitude of the coefficient reflects the strength of the association.

Feature	**Coefficient**
SeniorCitizen	0.24259186
tenure	−2.17919677
PhoneService	−0.48775946
OnlineSecurity	−0.37029458
OnlineBackup	−0.20221208
TechSupport	−0.28214541
PaperlessBilling	0.32610715
MonthlyCharges	0.79989955
InternetService_Fiber optic	0.73882175
InternetService_No	−0.63024013
Contract_Month-to-month	0.67989542
Contract_Two year	−0.46387532
PaymentMethod_Electronic check	0.39685993

For example, 'tenure' has a negative coefficient of -2.17919677, which suggests that longer tenure is associated with a lower probability of churn. On the other hand, 'MonthlyCharges' has a positive coefficient of 0.79989955, indicating that higher monthly charges are associated with a higher probability of churn.

- "SeniorCitizen": According to the positive coefficient, a senior citizen is more likely to experience customer churn. To solve this, you may think about putting in place focused retention measures for senior clients, such offering special discounts or specialized services.
- "SeniorCitizen": "tenure": The negative coefficient shows that tenure lengthens turnover likelihood. This suggests that you should prioritize increasing client loyalty and involvement. Customers that have supported your company for a long period may receive incentives or awards to honor their loyalty.
- "PhoneService": According to the negative coefficient, having phone service lowers the possibility of churn. Customers can be reminded of the advantages and worth of your phone service by emphasising its unique qualities.
- Both "OnlineSecurity" and "OnlineBackup" have negative coefficients, which indicates that having these services lowers the likelihood of churn. To persuade customers to keep their subscriptions, emphasize the value of data protection and the comfort these services offer.
- The negative correlation for "TechSupport" suggests that having access to technical support lowers churn. Make sure your technical support personnel is accessible and provide timely assistance to resolve any client difficulties or concerns.
- "PaperlessBilling": According to the positive coefficient, clients who receive paperless billing are more likely to leave their company. To encourage clients to transition to paperless billing, think about offering discounts or special deals.

- "MonthlyCharges": The positive coefficient implies that the likelihood of churn is increased by higher monthly costs. To keep clients who are price conscious, evaluate your pricing structure and think about providing competitive pricing or flexible plans.
- Customers having fiber optic internet service are more likely to churn than those who don't have it, according to the variables "InternetService Fiber optic" and "InternetService No". Examine the dependability and quality of your fiber optic service, and think about fixing any conceivable problems to increase client retention and happiness.
- The positive and negative coefficients for "Contract Month-to-month" and "Contract Two year" correspondingly show that consumers with month-to-month contracts are more likely to churn, whilst customers with two-year contracts are less likely to churn. By providing incentives or perks like discounted rates or extra services, you can persuade customers to sign contracts that are longer in duration.
- "PaymentMethod Electronic check": According to the positive coefficient, consumers who pay with electronic checks are more likely to leave. Encourage clients to use alternate payment options, which can increase convenience and lower churn rates, like automatic credit card payments or online payment gateways.

5.2 Conclusion

In conclusion, the logistic regression analysis aimed at predicting customer churn has provided valuable insights into the factors influencing customer retention. By examining the coefficients of the selected features, we have identified several key variables that significantly impact the likelihood of churn.

Based on our findings, the following recommendations can be made to reduce customer churn:

1. Implement targeted retention strategies for senior citizens to address their higher likelihood of churn.
2. Focus on enhancing customer loyalty and engagement, particularly for customers with longer tenure.
3. Emphasize the benefits and value of phone service to encourage customer retention.
4. Promote the importance of online security and backup services to reduce churn.
5. Ensure the availability of reliable and responsive technical support to address customer concerns.
6. Encourage customers to switch to paperless billing while offering incentives for the transition.
7. Evaluate pricing structures and consider offering competitive pricing or flexible plans to retain price-sensitive customers.
8. Assess the quality and reliability of fiber optic internet service and make improvements if necessary.
9. Encourage customers to sign longer-term contracts by offering incentives or benefits.
10. Encourage alternative payment methods to reduce churn among customers using electronic checks.

By implementing these recommendations, businesses can enhance customer satisfaction, loyalty, and ultimately reduce churn rates. It is important to continuously monitor customer behavior, adapt strategies based on changing market dynamics, and prioritize customer-centric approaches to maximize customer retention and long-term business success.

The references for this project: [1] ,[2], [3], [4] and [5]

References

[1] S. C. Bagley, H. White, and B. A. Golomb, "Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain," *Journal of clinical epidemiology*, vol. 54, no. 10, pp. 979–985, 2001.

[2] L. Niu, "A review of the application of logistic regression in educational research: Common issues, implications, and suggestions," *Educational Review*, vol. 72, no. 1, pp. 41–67, 2020.

[3] J. W. Osborne, "Bringing balance and technical accuracy to reporting odds ratios and the results of logistic regression analyses," *Best practices in quantitative methods*, vol. 11, no. 11, pp. 385–389, 2008.

[4] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, no. 1, pp. 3–14, 2002.

[5] C.-Y. J. Peng and T.-S. H. So, "Logistic regression analysis and reporting: A primer," *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, vol. 1, no. 1, pp. 31–70, 2002.