# **Analysing Customer Churn in the Telecom Industry**

## Objective

The objectives of this project are to develop machine learning models to forecast customer turnover, assess retention strategies, and study a telecom customer churn dataset to identify the variables impacting customer churn. To deliver useful insights and suggestions, we adhered to the data science lifecycle, which includes data exploration, hypothesis testing, A/B testing, and model construction.

## 1. Data Exploration and Preprocessing

### 1.1. Dataset Overview:

The dataset includes a number of variables pertaining to contract details, use trends, customer churn status, and customer demographics. Important characteristics consist of:   
  
**Customer demographics:** number of referrals, number of dependents, age, gender, and marital status.   
**Usage Information**: Internet Service, Average Monthly GB Download, Phone Service, and Average Monthly Long Distance Fees.   
**Details of the Contract:** Agreement, Monthly Fee, Total Fees, Total Refunds.   
Customer Status (Stayed, Churned, Joined) is the Churn Status.   
  
1.2. Managing Outliers and Missing Values:   
  
**Missing Values**: To ensure no appreciable loss of data integrity, rows with missing values can be amputated or removed. Z-scores and Interquartile Range (IQR) techniques are used to identify outliers, which are then dealt with by either eliminating or capping extreme results.

### 1.3. Feature Engineering:

**New Features:** In order to better understand consumer behavior, features like “Usage Ratios” and “Tenure Categories” were created.   
**Encoding:** One-hot encoding was used for categorical variables such as “Internet Service” and “Contract Type”.

### 1.4. Visualization:

Feature Distributions: The distributions of the main features were shown via boxplots and histograms.   
A chart with different colored rectangles

Description automatically generated

A chart with different colored squares

Description automatically generated

The first plot, a box plot comparing monthly charges across customer statuses, reveals that customers who stayed or churned have similar median monthly charges, suggesting that the monthly charge alone may not be a strong predictor of churn. However, new customers who joined tend to have lower monthly charges, possibly due to introductory offers. The second plot, a bar chart of customer status distribution, shows that the majority of customers have stayed with the service, while a significant number have churned, and a smaller portion has recently joined, highlighting the need for strategies to both retain existing customers and attract new ones.

2. Hypothesis Testing

### 2.1. Hypotheses Formulated:

**1. Contract Type vs Status of Customer:**   
H0: Customer churn is unaffected by contract type.   
H1: The kind of contract has a major impact on client attrition.   
  
**2. Customer Status vs. Online Security:**H0: Customer attrition is unaffected by online security.   
H1: Customer churn is greatly impacted by online security.

### 2.2. Statistical Tests:

**Chi-Square Test for Contract Type vs Customer Status:**

Chi2: 1233.75 p-value: 7.66e-266

Conclusion: Reject H0. Contract type significantly influences customer churn.

**Chi-Square Test for Online Security vs Customer Status:**

Chi2: 454.54 p-value: 1.98e-99

Conclusion: Reject H0. Online Security significantly influences customer churn.

### 2.3. Conclusions:

Two important variables influencing client turnover are contract type and online security. Clients that have unfavorable contract terms and inadequate internet security are more prone to leave.

## 3. A/B Testing

### 3.1. Design of A/B Test:

The purpose of A/B testing in this project is to evaluate the effectiveness of different strategies, such as retention campaigns or pricing changes, by comparing the outcomes between a control group (which does not receive the intervention) and a treatment group (which does). This helps determine if the strategy significantly reduces customer churn or improves other key metrics, providing data-driven insights for decision-making.

Objective: Evaluate the effectiveness of a discount offer to reduce churn.

Groups: Randomly assigned customers into control (no discount) and treatment (discount offer) groups.

### 3.2. Implementation:

**Control Group:** Received no special offers.

**Treatment Group:** Received a discount offer aimed at at-risk customers.

### 3.3. Results Analysis:

**Churn Rates:**

Control Group Churn Rate: 31.95%Treatment Group Churn Rate: 31.46%

**Statistical Test:**

Z-statistic: 0.1855 p-value: 0.8528

Conclusion: Fail to reject H0. The retention strategy did not significantly reduce churn.

## 4. Machine Learning Models

### 4.1. Data Preparation:

**Training and Testing Split:**

Data Splitting: To enable model training and assessment, the dataset was split into training and testing sets. With this divide, the model's performance can be verified on hypothetical data, providing a more accurate idea of how it would behave in actual situations.

### 4.2. Model Building and Evaluation:

#### 4. Gradient Boosting:

**Classification Report:**

|  |  |  |  |
| --- | --- | --- | --- |
| Precision | 0.73 (0) | 0.56 (1) | 0.86 (2) |
| Recall | 0.66 (0) | 0.54 (1) | 0.90 (2) |
| F1 Score | 0.69 (0) | 0.55 (1) | 0.88 (2) |
| Accuracy | **0.80** |  |  |
| ROC AUC Score | **0.9260** |  |  |
| MAE | **0.3547** |  |  |

The Gradient Boosting model stands out as the best-performing model among those tested. It achieves the highest accuracy at 80%, the best precision and recall for the churned class, and the highest overall ROC AUC score, indicating superior predictive capability. The lowest MAE further demonstrates its accuracy, making Gradient Boosting the most reliable model for predicting customer churn. This model offers the best combination of identifying who stays, churns, or joins, and minimizing prediction errors, making it highly suitable for the telecom dataset used in this analysis.

### 4.3. Model Optimization:

We used Grid Search, a methodical approach to determining the optimal hyperparameters for every machine learning model, to maximize model performance. Grid Search entails creating a grid of potential hyperparameter values and using cross-validation to assess the model's performance over this grid. The model is trained, verified, and its performance measures (such accuracy, precision, recall, or ROC AUC) are recorded for each set of hyperparameters. Grid Search finds the collection of hyperparameters that gives the model the highest performance by comparing these measures. By using this method, the models' accuracy and predictive capacity are increased when they are adjusted to their ideal parameters.

## 5. Key Insights and Recommendations

### 5.1. Key Insights:

* Factors Affecting Churn: Contract type and Online Security are major determinants of customer churn.
* Retention Strategy: The implemented discount offer did not have a significant impact on reducing churn.

### 5.2. Recommendations:

* Improve Contract Options: Introduce more flexible contract options to reduce churn rates.
* Enhance Online Security: Invest in improving online security features, as they significantly impact customer retention.
* Refine Retention Strategies: Develop and test new retention strategies beyond discounts to address customer needs more effectively.

## Conclusion

In summary, contract type and internet security are important factors influencing customer turnover in the telecom business. The Gradient Boosting model is the most successful at forecasting attrition because of its higher performance metrics. A/B testing revealed that a discount offer by itself was not sufficient to significantly lower turnover, indicating the need for more all-encompassing retention tactics. The telecom provider could strengthen online security features, offer more flexible contract alternatives, and look into retention tactics other than discounts to increase client retention. Predictive modeling and routine monitoring can help to improve these initiatives and increase customer happiness and engagement.