

FINAL REPORT

DATA OVERVIEW

The Telco Customer Churn Dataset, obtained from Kaggle (<https://www.kaggle.com/datasets/blatchar/telco-customer-churn/data>), contains information on 7,043 customers from a telecommunications company. The dataset includes customer demographics, service subscriptions, and payment details, with the binary response variable being Churn (Yes = 1, No = 0). Explanatory variables include both categorical (e.g., Partner, Internet Service, Payment Method) and continuous features (e.g., Tenure, Monthly Charges, Total Charges).

AIM AND RELEVANCE

The primary objective of this analysis is to predict customer churn using binary logistic regression and to identify significant factors associated with churn. Understanding customer churn is vital for telecommunications companies as it directly impacts revenue, growth, and the overall customer experience. By identifying significant churn drivers, companies can tailor retention strategies, optimize service offerings, and better manage customer relationships.

DATA EXPLORATION AND PREPARATION

The dataset was cleaned and scrutinized for missing values, outliers, and correlations. Key steps included:

- Addressing 11 missing values in the Total Charges column by replacing them with the median
- Converting categorical variables to factors and removing duplicate records.
- Discarding the Customer ID column to avoid multicollinearity, as it was deemed unnecessary.
- Checking for outliers in numerical variables (none was found).

DESCRIPTIVE ANALYSIS

The churn variable exhibited class imbalance, with 73% non-churners and 27% churners. Key numerical summaries:

- Tenure: Min = 0, Q1 = 9, Median = 29, Mean = 32.37, Q3 = 55, Max = 72
- Monthly Charges: Min = 18.25, Q1 = 35.50, Median = 70.35, Mean = 64.76, Q3 = 89.85, Max = 118.75
- Total Charges: Min = 18.8, Q1 = 402.2, Median = 1,397.5, Mean = 2,281.9, Q3 = 3,786.6, Max = 8,684.8

Categorical variables proportions:

- Gender: 50.5% Male, 49.5% Female
- Senior Citizen: 16% Yes, 84% No
- Online Backup: 43% No, 22% No internet service, 35% Yes
- Multiple Lines: 48% No, 10% No phone service, 42% Yes
- Internet Service: 34% DSL, 44% Fiber optic, 22% No
- Contract: 55% Month-to-month, 21% One year, 24% Two years
- Device Protection: 44% No, 22% No internet service, 34% Yes
- Streaming Movies: 40% No, 22% No internet service, 38% Yes
- Streaming TV: 40% No, 22% No internet service, 38% Yes

Additional categorical variable analysis was conducted but is not included here due to space. (See appendix for code.)

DATA VISUALIZATION

Visualizations were created for both categorical and numerical features (see appendix for graphs and code). Histograms and boxplots revealed:

- Total Charges: High frequency of values in the 0–2,000 range, followed by a sharp decline.
- Monthly Charges: Higher frequencies in the 0–20 and 80–100 ranges.
- Tenure: Frequency steadily declines as tenure increases.

Boxplots showed skewness in numerical variables, particularly Total Charges, which had a long upper whisker. Bar plots indicated more males than females, fewer senior citizens, and a higher proportion of customers without dependents, multiple lines, or phone service.

RELATIONSHIP BETWEEN CHURN AND OTHER VARIABLES

Numerical Variables: Churners tended to have higher monthly charges, lower total charges, and shorter tenure compared to non-churners. For example:

- Median Monthly charges: Churners = 80, Non-churners = 65.
- Median total charges: Churners = 1,400, Non-churners = 1,800.
- Median tenure: Churners = 10, Non-churners = 40.

Categorical Variables: Bar plots showed churn was more likely for customers with certain characteristics, such as month-to-month contracts, fiber optic internet, and paperless billing.

CORRELATION MATRIX

The correlation matrix revealed:

- Total Charges and Tenure: High correlation (0.8255).
- Monthly Charges and Total Charges: Moderate correlation (0.6509).
- Tenure and Monthly Charges: Low correlation (0.2479).

MODEL BUILDING

An initial logistic regression model (mod) was fit with an AIC of 5877.3. Significant variables included Senior Citizen status, Tenure, Multiple Lines, Internet Service, Contract, Paperless Billing, Payment Method, and Total Charges.

Polynomial Terms

Polynomial terms were tested for Tenure, Monthly Charges, and Total Charges (based on ggplot visualization), but were not significant and did not improve model fit (AIC unchanged). They were excluded for simplicity and interpretability.

Aliasing and Variance Inflation Factor (VIF)

Aliasing and VIF analysis revealed potential multicollinearity, particularly between Total Charges and Tenure. As a result, a reduced model (mod_red) was created, excluding highly correlated or insignificant variables such as Phone Service, Online Security, and Streaming Movies.

Interaction Terms

A third model (mod_int) was developed to test interaction terms. Significant predictors included Senior Citizen status, Tenure, Dependents, Multiple Lines, Internet Service, Contract, Payment Method, and Total Charges.

Variable Selection and Model Comparison

Stepwise selection was applied to all three models:

- Model 1: Lowest AIC (5870.9) but affected by multicollinearity (4 singularities).
- Model 2: No multicollinearity, slightly higher AIC (5929.1), better interpretability.
- Model 3: Best overall balance with AIC=5851.8, though some singularities remained.

The final model chosen included only significant predictors ($p < 0.05$), relevant interactions, and had the best AIC (5850.6). Key insights included:

- Fiber users churn 14.8 times more likely than DSL users.
- Two-year contracts reduce churn odds by 94%.
- Tenure's protective effect increases with contract length.

RESULTS AND RELEVANCE

Significant predictors from the final model:

- Tenure: Longer tenure significantly reduces churn ($p < 0.001$).
- Monthly Charges: Higher charges are linked to slightly lower churn risk ($p < 0.001$).
- Online Backup: Customers without internet service had lower churn; those with backup service had slightly higher churn ($p < 0.05$).
- Multiple Lines: Increased churn likelihood ($p < 0.001$); no phone service reduced churn.
- Internet Service = Fiber optic: Strongly associated with higher churn ($p < 0.001$).
- Device Protection = Yes: Reduced churn ($p < 0.001$).
- Streaming TV / Streaming Movies = Yes: Both increased churn risk ($p < 0.001$).
- Contract Type: One- and two-year contracts reduced churn ($p < 0.001$).

Interactions:

- Senior Citizen \times Payment Method: Certain combinations (e.g., non-seniors using mailed checks or credit cards) had lower churn ($p < 0.05$).
- Tenure \times Contract and Tenure \times Total Charges: These were significant, suggesting tenure's impact varies by contract type and spending.

Odds Ratio of Final Model

The odds ratios indicate how each variable influences the likelihood of churn. Longer tenure slightly decreases churn odds (OR = 0.95), and higher Monthly Charges also has a mild protective effect (OR = 0.93).

Service features such as having Online Backup (OR = 1.23), Device Protection (OR = 1.39), Streaming TV (OR = 2.66), and Streaming Movies (OR = 2.68) are associated with higher churn risk.

Customers with Multiple Lines were more likely to churn (OR = 1.96), while those with no phone service were less likely to (OR = 0.44). The largest effect was from Internet Service = Fiber optic, with

churn odds over 14 times higher ($OR = 14.06$). In contrast, one-year ($OR = 0.25$) and two-year contracts ($OR = 0.06$) greatly reduced churn.

Interaction terms such as tenure \times Contract suggest that tenure further lowers churn for contracted customers, and some Senior Citizen \times Payment Method combinations also reduced risk.

MODEL DIAGNOSTICS AND PERFORMANCE EVALUATION

- Hosmer-Lemeshow: p-value = 0.62 (chi-squared = 6.23, df = 8), indicating good model fit.
- The AIC = 5850.63 and BIC = 6008.40 indicating the model's overall quality, balancing fit and complexity. Suggests that the model fits the data well without being overly complex.

- **Confidence Intervals:**

- We are 95% confident that the true effect of tenure lies between -0.062 and -0.046, indicating that as tenure increases, the odds of the event (e.g., churn) decrease.
- We are 95% confident that the true effect of monthly charges lies between -0.092 and -0.049, suggesting that higher monthly charges are also associated with lower odds of the outcome.
- The odds ratio for tenure is 0.947, and we are 95% confident that the true odds ratio lies between 0.940 and 0.955, reinforcing the negative relationship with the outcome.
- We are 95% confident that having Fiber optic internet service increases the odds, with an effect between 2.10 and 3.19.
- For customers on a two-year contract, we are 95% confident the effect lies between -4.19 and -1.81, meaning such contracts strongly reduce the odds of the event.
- Interaction terms like tenure \times contract have small but significant positive effects, meaning that tenure slightly moderates the effect of contract type.

- **Hypothesis Testing:**

Chi-Square Test: The Chi-square statistic is 2345.519, and the p-value is 0, suggesting that the model is statistically significant overall.

ANOVA Results (Likelihood Ratio Test):

- Tenure, Online Backup, Monthly Charges, Multiple Lines, Internet Service, Streaming TV, Streaming Movies, Contract, Senior Citizen x Payment Method, Tenure x Contract, and Tenure x Total Charges all have p-values less than 0.05, indicating that they are significant predictors of the outcome variable.
- The p-value for Device Protection is 0.1357, which is greater than 0.05, suggesting that Device Protection is not a significant predictor in the model.

The results highlight that most of the predictors are significantly associated with the outcome, except for Device Protection. The model shows a strong fit, as evidenced by the low p-values for most variables.

- **Confusion Matrix and Statistics**

The confusion matrix results show that the model performs reasonably well, with an accuracy of 80.32%. The 95% confidence interval for accuracy is between 79.37% and 81.24%, indicating stable performance. The Kappa value is 0.4632, suggesting moderate agreement between the predicted and actual values. Key performance metrics include:

- Sensitivity (True Positive Rate): 89.99% – the model correctly identifies a high proportion of positive cases (Yes).
- Specificity (True Negative Rate): 53.56% – the model correctly identifies negative cases (No) at a lower rate.
- Positive Predictive Value (Precision): 84.29% – when the model predicts "Yes", it is correct 84.29% of the time.
- Negative Predictive Value: 65.90% – when the model predicts "No", it is correct 65.90% of the time.

The Balanced Accuracy is 71.77%, which accounts for both sensitivity and specificity. The model's performance is statistically significant, with a p-value for the McNemar's test and accuracy compared to the no information rate both being less than $2.2e-16$, suggesting a significant improvement over random guessing.

- **AUC and ROC curve:**

The ROC curve shows that the model effectively distinguishes between positive and negative cases, with a pronounced increase in the True Positive Rate (0.50 to 0.85). The blue line is above the red dotted line, indicating the model performs better than random guessing. The noticeable gap between the curves suggests a strong AUC, which reflects the model's excellent discriminatory power. A higher AUC closer to 1 indicates strong model performance.

KEY QUESTIONS AND ANSWERS

- Which factors most significantly impact churn?
 - Tenure, contract type, internet service, and payment method are key predictors.
- How do tenure, contract type, and payment method influence churn rates?
 - Longer tenure and longer contracts reduce churn. Certain payment methods (e.g., mailed checks for non-seniors) are associated with lower churn.
- Do interaction terms or alternative models enhance predictive accuracy?
 - Yes. Including interaction terms (e.g., tenure x contract) improved model performance and interpretability.

CONCLUSIONS

This analysis highlights that customer churn is influenced by service type, contract duration, tenure, and several features. The final logistics regression model, with interaction terms, effectively identifies customers at risk of churning and supports targeted retention strategies.

LIMITATIONS

- Model performance, though strong, could benefit from testing other algorithms like machine learning algorithms (e.g., random forests).
- Some features (e.g., device protection) were borderline significant and may warrant further exploration.

FUTURE ANALYSIS

- Explore ensemble models or deep learning approaches for comparison.
- Conduct time-series analysis if data over time is available.
- Implement cost-sensitive learning to prioritize high-value customers.
- Evaluate models using business KPIs (e.g., lifetime value, retention cost).

APPENDIX

Click on the HTML link (file:///Users/ololadeakinsanola/Desktop/STAT%20410%20-%20CATEGORICAL%20DATA/Project/Ololade_Akinsanola_Project.html) to access the code and graphs used in the analysis. You can also refer to the attached .qmd file on Sakai to run the code. Below are some of the key graphs used in this analysis.







