**Research Question**

What are the key factors influencing customer churn in the dataset, and how likely is a customer to churn based on specific characteristics (e.g., tenure, bandwidth usage, customer service calls)?

**Data Analysis Goal**

The main objectives of this data analysis are: Determine which customer features are the most significant predictors of churn. Use logistic regression to predict the probability that a specific customer will churn based on their characteristics. Provide actionable recommendations for the organization to reduce churn, such as improving service quality or offering targeted promotions to high-risk customers. Help the organization prioritize resources toward customers who are most at risk of churning, based on the results of the analysis. By understanding these factors, the organization can implement proactive strategies to retain customers and minimize churn rates, ultimately boosting profitability.

**Method Justification**

Logistic regression assumes that the dependent variable is binary. In this case, our target variable, churn, has two possible outcomes: either a customer churns (1) or does not churn (0). The observations should be independent of each other. Each data point represents a unique customer, and their behavior should not influence another customer in the data. Logistic regression does not require a linear relationship between the independent variables and the dependent variable itself. However, it assumes a linear relationship between the independent variables and the log odds of the dependent variable. The independent variables should not be perfectly correlated with one another. If two or more predictors are highly correlated, it may affect the stability of the model, leading to unreliable coefficient estimates. Python  provides powerful libraries (e.g., Pandas, Numpy,  and Matplotlib) to handle data cleaning, transformation, and visualization. This is useful for understanding the distribution of features, handling missing data, and identifying outliers before modeling. Python's scikit-learn library makes it easy to implement logistic regression models. Additionally, it provides robust tools for model evaluation to assess the performance of the logistic regression model and refine it through cross-validation or hyperparameter tuning.

Logistic regression is well-suited for the research question. The outcome of interest whether a customer will churn or not—is a binary variable, which fits the logistic regression model's assumption of a binary dependent variable. Logistic regression estimates the probability of an event occurring (in this case, customer churn). The coefficients from the model provide insights into how changes in independent variables affect the likelihood of a customer churning, making the results highly interpretable for decision-making. Logistic regression is computationally efficient and easy to implement, especially with large datasets, which is common in real-world business scenarios involving customer behavior. Logistic regression can help the organization determine which factors drive customer churn and create strategies to minimize churn by targeting at-risk customers.

**Data Preparation**

The data preparation process for logistic regression focused on ensuring the dataset was clean, transformed, and aligned with the research question of predicting customer churn. Since no missing values were present in the dataset, the focus was on converting the `Churn` variable to binary format (1 for "Yes" and 0 for "No"), a necessary step for logistic regression. Categorical variables such as `Contract` and `InternetService` were one-hot encoded to convert them into a numerical format suitable for modeling. Summary statistics were generated to provide an overview of both the dependent variable (`Churn`) and independent variables like `MonthlyCharge`, `Tenure`, and `Bandwidth\_GB\_Year`. Univariate and bivariate visualizations were created to explore the distributions of individual variables and their relationships with `Churn`. The data is now clean and prepared, making it ready for logistic regression analysis to assess the influence of various factors on customer churn.

A table of numbers and a few words

Description automatically generated with medium confidence

A close-up of a number

Description automatically generated

A graph of a graph of a graph

Description automatically generated with medium confidence

A graph of a diagram

Description automatically generated with medium confidence

A comparison of a graph

Description automatically generated

A graph of a graph of a graph

Description automatically generated with medium confidence

A graph of age and age

Description automatically generated

A graph of a graph of income and a bar of income

Description automatically generated with medium confidence

A diagram of a graph

Description automatically generated

A graph of a graph of a graph

Description automatically generated with medium confidence

A graph of contact information

Description automatically generated with medium confidence

A graph of a box plot

Description automatically generated with medium confidence

A graph of a graph of a graph

Description automatically generated with medium confidence

A chart with blue squares

Description automatically generated

A graph of a number of blue rectangular objects

Description automatically generated with medium confidence

A graph of blue rectangular bars

Description automatically generated with medium confidence

**Bivariate Visualizations/Statistics**

**A graph of blue and orange bars

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph of blue and orange bars

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A chart with blue and orange squares

Description automatically generated**

**A screenshot of a number chart

Description automatically generated**

**A diagram of a diagram

Description automatically generated with medium confidence**

A screenshot of a computer

Description automatically generated

A chart with different colored squares

Description automatically generated

A screenshot of a number

Description automatically generated

A graph of blue and orange squares

Description automatically generated

A table with numbers and letters

Description automatically generated

A chart with blue squares and orange squares

Description automatically generated

A table with numbers and symbols

Description automatically generated

A blue and orange squares

Description automatically generated

A screenshot of a number

Description automatically generated

A chart with blue squares and black lines

Description automatically generated

A screenshot of a computer program

Description automatically generated

A graph with blue squares and orange squares

Description automatically generated

A screenshot of a number

Description automatically generated

A white background with black and white clouds

Description automatically generated

A chart with blue squares and orange squares

Description automatically generated

A number and numbers on a white background

Description automatically generated

A blue and orange rectangular boxes

Description automatically generated

A white background with black text

Description automatically generated

A diagram of a number of blue squares

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated

**Modeling Statistics**

A screenshot of a computer screen

Description automatically generated

**Initial Logistic Regression Model**

**A screenshot of a computer

Description automatically generated**

**Reduced Logistic Regression Model**

**A screenshot of a computer

Description automatically generated**

**A blue squares with white text

Description automatically generated**

**Model Comparison and Analysis**

The two logistic regression models aim to predict customer churn using different sets of features. The first model included 15 features, while the second model reduced this to 7, with a slight improvement in accuracy from 89.65% to 89.8%. The key difference between the models is the removal of statistically insignificant variables, such as Population, Children, Income, Outage\_sec\_perweek, Email, Contacts, and Yearly\_equip\_failure. These features were removed based on their p-values, which measure the statistical significance of each feature. In the first model, many of these features had p-values greater than 0.05, indicating that their contribution to predicting churn was not statistically significant. After each feature removal, the model is refitted with the remaining variables, which recalculates the p-values for the remaining features. This process allows us to detect whether the removal of one feature affects the p-values of the others. When no feature has a p-value above 0.05 we are left with only statistically significant variables.

By eliminating features with high p-values, the second model reduces complexity without negatively affecting accuracy, focusing on the most predictive variables such as MonthlyCharge, Bandwidth\_GB\_Year, Tenure, and Contract Type. This process of feature elimination based on p-value ensures that only the most relevant variables are retained, improving model interpretability and avoiding overfitting while maintaining or improving model performance.

The confusion matrix shows the model's performance in predicting customer churn. It correctly predicted 1364 customers who did not churn (true negatives) and 432 customers who did churn (true positives). However, it incorrectly predicted 92 customers would churn when they did not (false positives) and 112 customers would not churn when they actually did (false negatives). This indicates that the model generally performs well, accurately identifying the majority of both churned and non-churned customers, but it still makes some errors, particularly in missing a small number of actual churners.

**Data Summary and Implications**

The reduced logistic regression model identified several significant factors influencing customer churn, resulting in the following equation:

log ( P(Churn = 1) / 1 - P(Churn = 1) ) = -6.6401 + 0.0485 x MonthlyCharge + 0.0015 x Bandwidth\_GB\_Year + 0.0060 x Age - 0.2315 x Tenure - 3.2708 x Contract\_One\_year - 3.2608 x Contract\_Two\_Year - 1.5619 x InternetService\_Fiber\_Optic

The coefficients indicate that higher MonthlyCharge (0.0485) and greater Bandwidth\_GB\_Year (0.0015) are associated with a higher likelihood of churn, while longer Tenure (-0.2315) and having a One-year contract (-3.2708), Two-year contract (-3.2608), or using Fiber-optic internet (-1.5619) significantly reduce the likelihood of churn. Age (0.0060) has a small but positive effect on churn, meaning older customers are slightly more likely to churn. Statistically, all these features are highly significant (p-values < 0.05). The model’s Pseudo R-squared is 0.5919, indicating that the model provides a substantial improvement over the null model. It’s important to note that Pseudo R-squared does not represent explained variance in the same way as R-squared in linear regression. Looking at our continuous variables exponentiating the coefficient helps us interpret the change in the odds ratio of the dependent variable for each one-unit increase in the independent variable.

The coefficient for MonthlyCharge is 0.0485. When exponentiated exp(0.0485) ≈ 1.0496 this means that for each $1 increase in MonthlyCharge, the relative chance of a customer churning increases by 4.97%. In other words, customers have a 4.97% larger relative chance of churn for each additional dollar in charges, assuming all other variables are held constant. The coefficient for Tenure is -0.2315. When exponentiated exp(−0.2315) ≈ 0.7933 this means that for each additional month of tenure, the relative chance of churn decreases by 20.66%. In other words, for each extra month a customer stays with the company, they have a 20.66% smaller relative chance of churn, assuming all other factors remain constant. The same concept follows for the remaining continuous variables, for each additional gigabyte of Bandwidth\_GB\_Year exp(0.0015) ≈ 1.0015 used annually, the relative chance of churn increases by 0.15%. This means that customers using more bandwidth have a 0.15% larger relative chance of churn for each additional gigabyte, assuming all other factors remain constant.

Age, exp(0.0060) ≈ 1.006 means for each additional year of age, the relative chance of churn increases by 0.60%. Older customers have a 0.60% larger relative chance of churn for every additional year in age. Looking at our binary independent variables, the interpretation compares the odds of an event happening when the variable is "on" (value is 1) versus when it is "off" (value is 0). We compare the category in the model to the baseline reference category. Contract\_One\_Year has a coefficient of -3.2708 when exponentiated exp(-3.2708) ≈ 0.0378. Meaning that customers on a one-year contract have a 96.22% less relative chance of churn than those in the baseline category (month-to-month).

Similarly, Contract\_Two\_Year has a coefficient of -3.2608. When exponentiated exp(−3.2608) ≈ 0.0382. This means customers on a two-year contract have a 96.18% less relative chance of churn compared to those of the baseline (month-to-month). InternetService\_Fiber\_Optic has a coefficient of -1.5619, exponentiated exp(−1.5619) ≈ 0.2099. Customers with InternetService\_Fiber\_Optic have a 79.01% smaller relative chance of churn when compared to DSL customers and customers with no internet services, so in summary for our continuous variables we obtain the odds ratio, which tells us how much the odds of the outcome (customer churn) increase or decrease for each one-unit increase in that independent variable and for our binary variables the coefficient gives us an odds ratio that compares the odds of the outcome between the "on" and "off" groups.

**Practical Significance**

MonthlyCharge indicates that higher costs slightly increase churn risk, to combat this the company could offer personalized discounts or added value for high-paying customers to maintain customer loyalty. Bandwidth\_GB\_Year has a minimal impact on churn, suggesting that high-usage customers aren’t necessarily at risk of leaving, but targeted service packages could help increase satisfaction among heavy users to promote less churn. Age also has a minor effect, with older customers being slightly more likely to churn. Possibly implying that older customers may benefit from simpler services, loyalty programs, or targeted customer support to feel valued and reduce their likelihood of churning. Tenure shows a strong effect, where longer-tenured customers are far less likely to churn. The takeaway is that investing in retention strategies early in the customer journey with onboarding programs, loyalty incentives, and ongoing engagement can significantly reduce churn over time.

Contract\_One\_Year and Contract\_Two\_Year both show that customers on longer contracts are significantly less likely to leave, with over 96% lower churn odds compared to month-to-month customers. This suggests that promoting these contracts through benefits like discounts, special perks, or bundled offerings could help lock in customer loyalty. Long-term contracts provide stability, reducing the churn risk and creating a predictable revenue stream for the company. InternetService\_Fiber\_Optic has a substantial impact in reducing churn, expanding fiber-optic infrastructure and marketing its advantages to customers such as faster speeds and better reliability could significantly improve customer satisfaction and reduce churn rates. The overall practical significance of these variables show that offering attractive long-term contracts, expanding premium services like fiber-optic internet, and focusing on value-based engagement are key strategies for reducing customer churn and enhancing customer loyalty.

**Limitations**

The logistic regression model has several limitations that impact its reliability and accuracy. Multicollinearity is a concern, particularly between variables like Tenure and Contract Type, which may influence each other and distort the true effect of each predictor. The model also assumes a linear relationship between the independent variables and the log odds of churn, which may oversimplify complex real-world relationships. Additionally, the assumption of independence of observations may not hold, as customers in similar geographic or social groups could influence each other's behavior. The model also suffers from limited feature representation, as important predictors like customer satisfaction or service issues are not included, potentially reducing its predictive power. Lastly, the interpretation of odds ratios can be challenging, as they represent the change in odds rather than direct probabilities. This can make it difficult for non-technical stakeholders to fully understand the practical implications of the model’s findings.

**Recommendations**

Since long-term contracts significantly reduce churn, the company should encourage customers to adopt these contracts by offering incentives such as discounts or added benefits. Target higher monthly charges, customers with higher monthly charges are more likely to churn, so targeted discounts or personalized plans could help retain this segment of customers. New customers with shorter tenures are more prone to churn, so early engagement and satisfaction programs could prevent this. Monitor high bandwidth users, although the effect of bandwidth usage on churn is small, high-usage customers could benefit from tailored packages, to enhance retention. Improve internet service offerings: Fiber-optic users are less likely to churn, suggesting that expanding this service or improving other internet service options could boost customer satisfaction and reduce churn. By focusing on these key drivers of churn, the company can implement effective retention strategies to reduce customer loss.

**References**

geeksforgeeks.org accessed 4 October 2024, https://www.geeksforgeeks.org/numpy-exp-python/

scikit-learn.org accessed 22 September 2024, https://scikit-learn.org/stable/modules/feature\_selection.html

scikit-learn.org, accessed 22 September 2024, https://scikit-learn.org/stable/auto\_examples/model\_selection/plot\_confusion\_matrix.html

scikit-learn.org, accessed 22 September 2024, https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html

groverjatin.medium.com, accessed 06 August 2024, https://groverjatin.medium.com/short-python-code-for-backward-elimination-with-detailed-explanation-52894a9a7880

scikit-learn.org, accessed 06 August 2024, https://scikit learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html

www.statology.org, accessed 06 August 2024, https://www.statology.org/standardized-residuals-python/

sphweb.bumc.bu.edu, accessed 06 August 2024, https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713\_multivariablemethods/bs704-ep713\_multivariablemethods2.html

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy-1.6.3/reference/generated/scipy.stats.ttest\_ind.html.

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html#skew.

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kurtosis.html#kurtosis.

SciPy.org, accessed 05 July 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html#scipy.stats.chisquare.

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA) accessed 09 June 2024, https://apps.cgp-oex.wgu.edu.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

W3schools.com accessed, 09 June 2024, https://www.w3schools.com/python/numpy/numpy\_array\_iterating.asp.

W3schools.com accessed, 09 June 2024, https://www.w3schools.com/python/pandas/default.asp.