Project Report: Multiple Linear Regression Analysis on Bandwidth Usage

1. Introduction

The primary objective of this project was to analyze the factors influencing the yearly bandwidth usage (measured in GB) of customers using multiple linear regression. The analysis was conducted on a customer churn dataset, where the target variable was **`Bandwidth_GB_Year`**.

2. Data Description

The dataset, **`churn_clean.csv`**, contains various demographic and service-related variables. The variables of interest for this analysis included:

- **Demographic Variables:** `Area`, `Children`, `Age`, `Income`, `Marital`, `Gender`
- **Service Variables:** `Churn`, `Outage_sec_perweek`, `Yearly_equip_failure`, `Tenure`, `MonthlyCharge`, `Bandwidth_GB_Year` (dependent variable)

3. Data Preprocessing

3.1 Variable Selection

Only relevant columns were selected for the analysis.

3.2 Data Cleaning

The dataset was checked for missing values and duplicate records. No missing values or duplicates were found.

3.3 Summary Statistics

Summary statistics were computed for both continuous and categorical variables to understand the data distribution.

4. Exploratory Data Analysis (EDA)

4.1 Histograms

Continuous variables like `Bandwidth_GB_Year`, `Children`, `Age`, `Income`, `Outage_sec_perweek`, `Yearly_equip_failure`, `Tenure`, and `MonthlyCharge` were plotted to visualize their distributions.

4.2 Countplots

Categorical variables (`Area`, `Marital`, `Gender`, `Churn`) were visualized using countplots.

4.3 Bivariate Analysis

4.3.1 Regression Plots

Relationships between `Bandwidth_GB_Year` and continuous variables (`Children`, `Age`, `Income`, etc.) were analyzed.

4.3.2 Boxplots

The impact of categorical variables on `Bandwidth_GB_Year` was assessed using boxplots.

5. Modeling

5.1 Initial Model

A multiple linear regression model was built using all selected predictors. The model summary indicated the statistical significance of the predictors.

5.2 Model Evaluation

The initial model had an **MSE (Mean Squared Error)** of `mse_all` and **RSE (Residual Standard Error)** of `rse_all`. R-squared and Adjusted R-squared were computed to assess the model's explanatory power.

5.3 Variance Inflation Factor (VIF) Analysis

Multicollinearity was evaluated using VIF. Two variables with high VIF (`Outage_sec_perweek` and `MonthlyCharge`) were removed from the model.

5.4 Reduced Models

The model was refined by dropping variables with high VIF, leading to improved model performance. The final reduced model included the predictors `Children`, `Age`, `Marital_cat`, `Gender_cat`, `Churn_cat`, and `Tenure`.

5.5 Final Model Summary

The final model showed an **MSE** of 60923.13862410184, an **RSE** of 246.82613035110737 and R-squared values that indicated a reasonable fit.

6. Results

- **Key Predictors:** `Children`, `Age`, `Tenure`, `Marital status`, `Gender`, and `Churn` were found to be significant predictors of bandwidth usage.
- **Model Performance:** The reduced model was more parsimonious and performed comparably to the initial model in terms of explanatory power.

7. Conclusion

This analysis successfully identified key factors influencing customer bandwidth usage. The final model, which excluded variables with high multicollinearity, provided a robust explanation for variations in bandwidth usage across different customer demographics and service-related factors. The results could be used to inform customer retention strategies and optimize service offerings.

8. Future Work

Future analyses could explore non-linear relationships, interactions between variables, or apply machine learning techniques for potentially improved predictions. Additionally, segmenting the customer base might reveal more tailored insights for specific groups.