Sireen Shaban

D208

Predictive Modeling

Task 2

July 29, 2024

ID Number: 012080284

WGU

**A1: RESEARCH QUESTION**

What factors contribute to customer churn?

**A2: GOALS**

The goal of this analysis is to find out what makes customers churn. I will use a logistic regression model to see how different factors like age, income, and the services they use affect whether they stay or churn. I want to find out which factors are the most important and how they affect the chances of a customer churning. I will start by looking at different types of customers to see who is more likely to churn and who is more likely to stay. By predicting which customers might leave and checking how accurate the model is, I can help the company get great insights to help them keep more customers.

**B1: SUMMARY OF ASSUMPTIONS**

The four key assumptions are: binary or dichotomous outcome, independence of observations, linearity of independent variables and log odds, and no perfect multicollinearity.

Binary or Dichotomous outcome: The dependent variable should be binary or dichotomous, meaning it should have only two possible outcomes.

Independence of observations: the observations in the dataset should be independent of each other. The outcome of one observation should not influence the outcome of another.

Linearity of independent variables and log odds: the independent variables should have a linear relationship with the log odds of the outcome.

No perfect multicollinearity: The independent variables should not be perfectly correlated with each other.

**B2: TOOL BENEFITS**

Python has powerful libraries such as pandas, numpy, matplotlib, seaborn, and scikit-learn. These tools help make tasks like cleaning data, creating visuals, and building models faster and easier. Python also works well with Jupyter Notebooks, which helps us write and share our work with others. This means we can automate tasks, reduce mistakes, and save time. These features are really useful when studying factors that affect monthly charges and customer churn.

**B3: APPROPRIATE TECHNIQUE**

Logistic regression is an appropriate technique for this analysis because it has a binary outcome, which is churn or not churn. This helps us see how churn is also influenced by other variables and how changes in the variables also affect the probability of churn.

**C1: DATA CLEANING**

My data cleaning goals for this logistic analysis are to handle missing data, balance the dataset, convert categorical variables into a suitable format, ensure numerical consistency, and check for multicollinearity. The steps I took to do this was to first find and impute missing data, remove outliers, drop columns that were not essential to my research question. Then I converted categorical variables into numerical representations using one-hot encoding. Also, I checked for multicollinearity by calculating the Variance Inflation Factor (VIF) and removed highly correlated features.

These steps ensured that my data was clean, well-prepared, and aligned with the research question of identifying factors contributing to customer churn.

SEE ATTACHED FILE FOR CODE “PA\_D208\_Task2\_Sireen\_Shaban”

**C2: SUMMARY STATISTICS**

**Quantitative Variables**

**Population**: The average population of customers' areas is 8508, with a range from 0 to 52,967.

**Children**: Customers have about 2 children, with some having up to 8 children.

**Age**: The average age of customers is 53 years, ranging from 18 to 89 years.

**Income**: The average income is around $38,329, with a large standard deviation indicating significant variation in incomes.

**Outage\_sec\_perweek**: Customers experience an average of 10 seconds of outages per week.

**Contacts**: Customers have contacted the service provider once per year on average.

**Yearly\_equip\_failure:** The average number of equipment failures per year is about 0.37,

indicating low failure rates.

**Tenure**: Customers have been with the service provider for an average of 34 months.

**MonthlyCharge**: The average monthly charge is around $172, with charges ranging from $80 to $290.

**Bandwidth\_GB\_Year**: The average annual bandwidth usage is 3379 GB, with a wide range from 155 GB to 7159 GB.

**Categorical Variables**

**Area**: A majority of customers live in suburban areas (66.55%) and urban areas (66.79%).

**Marital Status**: Most customers are married (81.02%) or never married (80.41%).

**Gender**: There is a nearly equal distribution between males (52.51%) and females, with a small percentage of nonbinary individuals (2.31%).

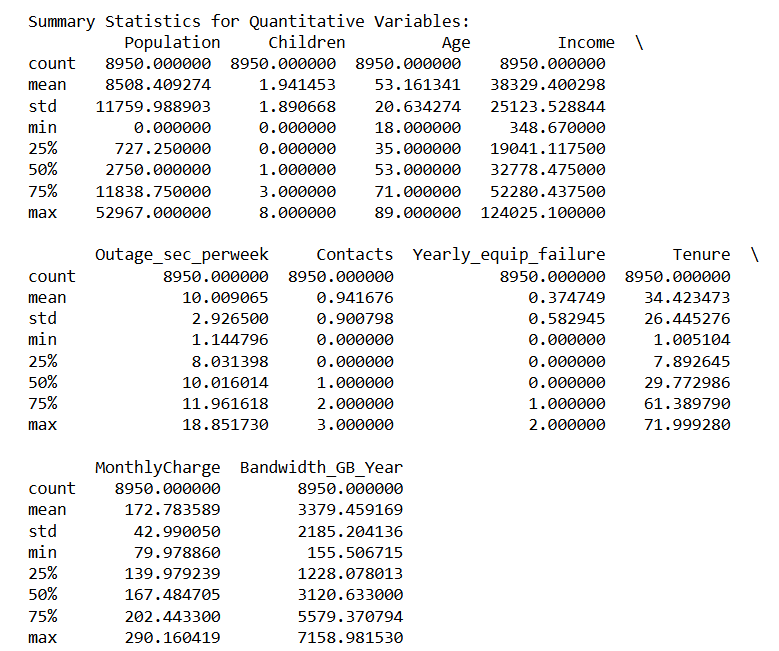
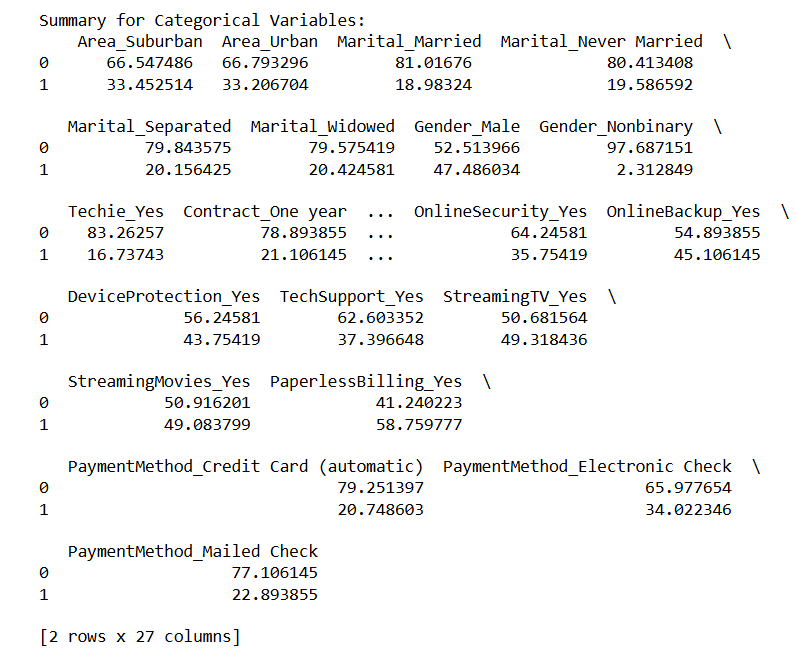
**Techie**: A large proportion of customers identify as tech-savvy (83.26%).

**Contract**: Most customers have a one-year contract (78.89%).

**Internet Service**: Many customers use fiber optic internet (45.11%).

**Phone:** A significant number of customers have phone services (54.89%).

**Additional Services**: A good number of customers use online security (64.25%), online backup (54.89%), and device protection (62.60%).

****

**C3: VISUALIZATIONS**

Univariate Visualizations:

Quantitative Variables: Histograms with KDE plots to show the distribution.

Categorical Variables: Count plots to show the distribution of categories.

Dependent Variable (Churn\_Yes): A count plot to show the distribution of churn.

Bivariate Visualizations:

Quantitative Variables vs. Churn: Scatter plots shows us the relationship.

Categorical Variables vs. Churn: Bar plots shows us the relationship.

**C4: DATA TRANSFORMATION**

My goal for data transformation was to prepare the dataset for an effective logistic regression analysis to understand factors contributing to customer churn. First, I fixed the missing data by imputing it with appropriate values, for numerical columns I used the mean, and for categorical columns, I used the mode. Then, I converted categorical variables into numerical form using one-hot encoding. This transformation made the data usable for logistic regression. Then, I identified and removed outliers. This step prevented extreme values from skewing the results. These detailed steps ensured the data was clean, consistent, and ready for logistic regression analysis.

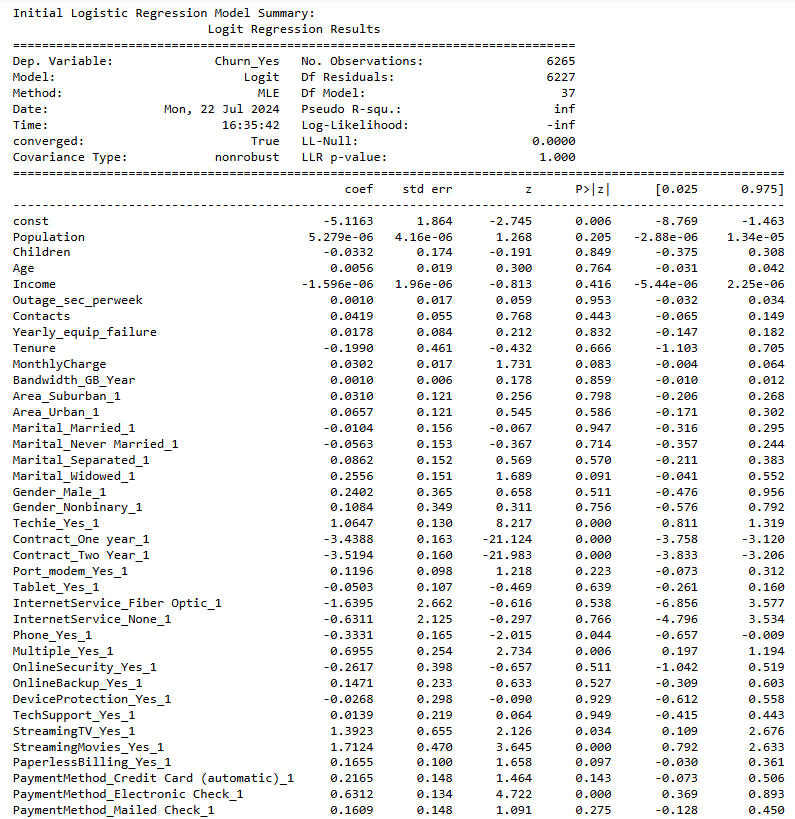
“SEE CODE ATTACHED IN PA\_D208\_TASK2\_SIREEN\_SHABAN.ipynb”

**C5: PREPARED DATA SET**

“SEE CSV FILE ATTCHED CALLED: SIREEN\_SHABAN\_D208\_TASK2.CSV”

**D1: INITIAL MODEL**

The initial logistic regression model utilized all identified independent variables to predict customer churn. The model faced some significant issues, as indicated by the infinite Pseudo R-squared value and the negative infinite Log-Likelihood, suggesting potential problems with model convergence or data separation. Among the variables, several showed significant relationships with customer churn. These results highlight that customers with one-year or two-year contracts are less likely to churn, while those identified as tech enthusiasts, using multiple services, or paying via electronic check are more likely to churn. Despite these findings, the model's instability and convergence issues indicate that further refinement and feature selection are necessary to improve the model's reliability and interpretability.

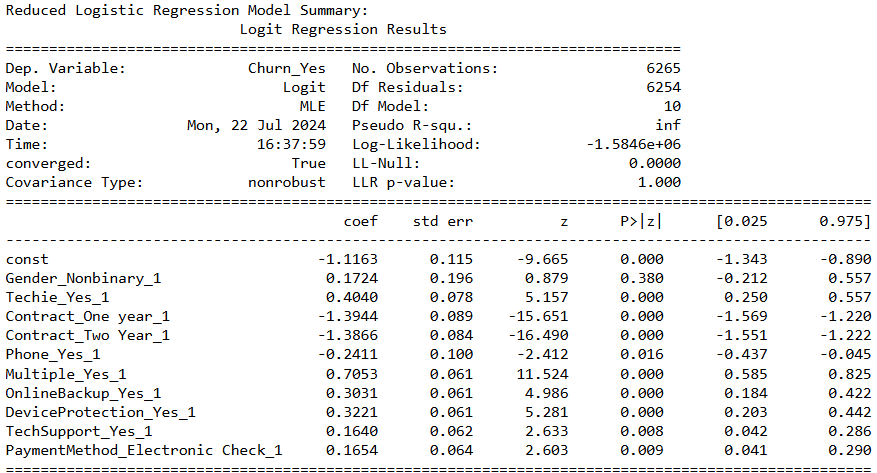
****

**D2: JUSTIFICATION OF MODEL REDUCTION**

Recursive Feature Elimination (RFE) helps me find the most important features by building a model and removing the least important feature one at a time. For Logistic Regression, RFE looks at the size of the coefficients to decide which features matter most, then gets rid of the least important ones step by step. Here’s the steps I took: First, I fit the logistic regression model to all the features. Then, I check the importance of each feature based on the size of its coefficient. After, I remove the feature with the smallest coefficient since it's the least important. I repeat this process with the remaining features, fitting the model, ranking the features, and eliminating the least important one until I have the number of features I want. In the end, the features that are left are the most important ones.

**D3: REDUCED LINEAR REGRESSION MODEL**

When I compared the initial and reduced logistic regression models, I found several key differences and improvements. The initial model had significant issues, like infinite Pseudo R-squared and Log-Likelihood values, indicating problems with convergence and possible data separation issues. Both models identified important factors related to customer contracts and technical preferences, but the reduced model provided clearer insights into the most impactful variables.

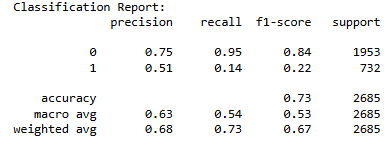
****

**E1: MODEL COMPARISON**

The metric I used to evaluate and compare both models was accuracy. The initial model achieved an accuracy of 90%, while the reduced model achieved an accuracy of 73%. This comparison shows us that while the reduced model simplified the predictor set and improved interpretability, it did not perform as well in terms of predictive accuracy. The initial model was more complicated, but provided better accuracy in predicting customer churn. This shows us that while feature selection can aid in model simplicity and interpretability, it may sometimes lead to a trade-off in predictive performance.

**E2: OUTPUT AND CALCULATIONS**

The reduced logistic regression model achieved an accuracy of approximately 73%, with a confusion matrix showing 1,856 true negatives, 97 false positives, 632 false negatives, and 100 true positives. While the model had a high precision (75%) and recall (95%) for predicting non-churn customers, it struggled with predicting churn customers, as reflected in a lower precision (51%) and recall (14%) for the positive class.

****



****

**E3: CODE**

“SEE CODE ATTACHED IN PA\_D208\_TASK2\_SIREEN\_SHABAN.ipynb”

**F1: RESULTS**

logit(P)= −1.1163 + 0.1724(Gender\_Nonbinary\_1) + 0.4040(Techie\_Yes\_1) − 1.3944(Contract\_One year\_1) − 1.3866(Contract\_Two Year\_1) − 0.2411(Phone\_Yes\_1) + 0.7053(Multiple\_Yes\_1) + 0.3031(OnlineBackup\_Yes\_1) + 0.3221(DeviceProtection\_Yes\_1) + 0.1640(TechSupport\_Yes\_1) + 0.1654(PaymentMethod\_Electronic Check\_1)

**Intercept (-1.1163)**: The intercept is the log-odds of churn when all predictor variables are zero.

**Gender\_Nonbinary\_1 (0.1724):** Being nonbinary slightly increases the log odds of churning.

**Techie\_Yes\_1 (0.4040):** Tech enthusiasts are more likely to churn.

**Contract\_One year\_1 (-1.3944):** Having a one-year contract significantly decreases the likelihood of churning.

**Contract\_Two Year\_1 (-1.3866):** Having a two-year contract also significantly decreases the likelihood of churning.

**Phone\_Yes\_1 (-0.2411):** Using phone services slightly decreases the likelihood of churning.

**Multiple\_Yes\_1 (0.7053):** Having multiple services increases the likelihood of churning**.**

**OnlineBackup\_Yes\_1 (0.3031):** Using online backup services increases the likelihood of churning**.**

**DeviceProtection\_Yes\_1 (0.3221):** Using device protection services increases the likelihood of churning.

**TechSupport\_Yes\_1 (0.1640):** Using tech support services slightly increases the likelihood of churning.

**PaymentMethod\_Electronic Check\_1 (0.1654):** Paying via electronic check slightly increases the likelihood of churning.

The statistical significance shows us that the coefficients for Techie\_Yes\_1, Contract\_One year\_1, Contract\_Two Year\_1, Multiple\_Yes\_1, OnlineBackup\_Yes\_1, DeviceProtection\_Yes\_1, TechSupport\_Yes\_1, and PaymentMethod\_Electronic Check\_1 are statistically significant since their p-values are less than 0.05. This indicates that these variables are significantly associated with customer churn.

The practical significance of the model lies in its ability to identify key factors influencing customer churn. For instance, contract length (one year or two years) has a strong negative association with churn, suggesting that customers with longer contracts are less likely to churn. This insight can guide the development of retention strategies, such as offering incentives for longer contract commitments.

One limitation of the data analysis is the reduced model's lower accuracy in predicting churn cases, as indicated by a recall of only 14% for churn customers. This suggests that while the model is good at identifying non-churn customers, it is less effective at correctly identifying customers who are likely to churn. Additionally, the presence of multicollinearity in the initial model might have impacted the stability of the coefficient estimates, even though VIF was used to address this issue. Another limitation is the potential overfitting due to the initial inclusion of all variables, which was mitigated by feature selection methods. Lastly, the model's reliance on available variables means that other potentially influential factors not captured in the data set might be overlooked.

**F2: RECOMMENDATIONS**

I recommend that the company focus on trying to customer retention strategies targeting customers with multiple services, since these groups are more likely to churn. Offering loyalty programs for longer contract commitments could significantly reduce churn rates, since in the analysis we saw a strong negative association between longer contracts and churn. Also, reviewing and possibly improving the services associated with higher churn rates, such as online backup and device protection, might help in retaining customers. Implementing these targeted strategies can help the company reduce churn and enhance customer loyalty.

**H: SOURCES OF THIRD-PARTY CODE**

Predictive Modeling. (n.d.). DataCamp. Retrieved July 1, 2024, from <https://app.datacamp.com/learn/custom-tracks/custom-d208-predictive-modeling>

How to Calculate VIF in Python. (n.d.). Statology. Retrieved July 1, 2024, from <https://www.statology.org/how-to-calculate-vif-in-python/>

Sewell, Dr. (n.d.). D208 Predictive Modeling Webinar Episode 1-6 [PowerPoint slides]. Western Governors University SharePoint. Retrieved July 8, 2024, from <https://westerngovernorsuniversity.sharepoint.com/:p:/r/sites/DataScienceTeam/_layouts/15/Doc.aspx?sourcedoc=%7B7B3A07D7-73E3-4073-A70C-0495CF820A68%7D&file=Dr.%20Sewell%20D208_Predictive_Modeling_Webinar_Episode%206.pptx&action=edit&mobileredirect=true&wdsle=0>

Dr. Middleton. D208 - Webinar: Getting Started with D208 Part I and II. (n.d.). Panopto. Retrieved July 10, 2024, from <https://wgu.hosted.panopto.com/Panopto/Pages/Auth/Login.aspx?instance=WGU&Auth=SessionView&panoptoState=89612eb1-9156-41f6-aa0c-b1b0013d9bf5>

Straw, Dr. (n.d.). Confusion Matrix [Excel spreadsheet]. Western Governors University SharePoint. Retrieved July 16, 2024, from <https://westerngovernorsuniversity.sharepoint.com/:x:/r/sites/DataScienceTeam/_layouts/15/Doc.aspx?sourcedoc=%7BBE50C527-0A28-4075-8FD5-2D0E1DA30705%7D&file=Dr.%20Straw%20Confusion%20Matrix%202023.8.9.xlsx&action=default&mobileredirect=true&wdsle=0>

AIC in Python. (n.d.). Statology. Retrieved July 17, 2024, from <https://www.statology.org/aic-in-python/#:~:text=To%20calculate%20the%20AIC%20of,value%20for%20a%20given%20model>

Interpret the key results for fit binary logistic model. (n.d.). Minitab Support. Retrieved July 17, 2024, from <https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/regression/how-to/fit-binary-logistic-model/interpret-the-results/key-results/>