Telecom Customer Churn Prediction

Final Project Work for
MSA 8150: Machine Learning for Analytics

April 23, 2024

Author: Surya Vegesna

Table of Contents

Introduction:

What is Customer Churn?

Customer churn is defined as when customers or subscribers discontinue doing business with a firm or service.

Customers in the telecom industry can choose from a variety of service providers and actively switch from one to the next. The telecommunications business has an annual churn rate of 15-25 percent in this highly competitive market.

To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

To detect early signs of potential churn, one must first develop a holistic view of the customers and their interactions across numerous channels, including store/branch visits, product purchase histories, customer service calls, Web-based transactions, and social media interactions, to mention a few.

As a result, by addressing churn, these businesses may not only preserve their market position, but also grow and thrive. More customers they have in their network, the lower the cost of initiation and the larger the profit. As a result, the company's key focus for success is reducing client attrition and implementing effective retention strategy.

Problem Statement

A telecommunications company is experiencing high customer churn rates, leading to revenue loss and decreased customer satisfaction. Identifying the factors that contribute to customer churn and building an accurate predictive model to forecast churn risk can help the company proactively retain valuable customers and reduce churn.

Proposed Solution

- To address the issue of high customer churn, we propose building a predictive model using machine learning techniques. The goal is to predict whether a customer is likely to churn based on various features such as contract type, tenure, monthly charges, internet service type, and customer demographics.

Data Description

- The initial dataset consists of 7043 rows and 21 columns
- The dataset includes the following columns:

  'customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn'.

Description of Columns:

- **Churn** -Customers who left within the last month
- **Online Services that each customer has signed up for** – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- **Customer account information** - how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- **Demographic info about customers** – gender, age range, Partner and Dependents.

Data Cleaning:

- Identified that the 'TotalCharges' column is in object datatype, which should be numeric for proper analysis. Converted 'TotalCharges' to numeric type using the pd.to_numeric() function, setting errors='coerce' to convert non-convertible values to NaN.
- Upon inspection, it was found that there are 11 missing values in the 'Total Charges' column.
- To handle missing values, particularly for new customers with a tenure of 0, 'TotalCharges' was set equal to 'MonthlyCharges', assuming that their first bill might be equivalent to one month of service.

Given that the Tenure column is 0 for the records where TotalCharges is missing, and since MonthlyCharges is not empty for these records, it seems these customers are new and have not yet been billed beyond the first month. we set TotalCharges equal to MonthlyCharges for these records. This assumes that their first bill might be equivalent to one month of service.

```
customerID          object
gender              object
SeniorCitizen        int64
Partner             object
Dependents          object
tenure               int64
PhoneService        object
MultipleLines       object
InternetService     object
OnlineSecurity      object
OnlineBackup        object
DeviceProtection    object
TechSupport         object
StreamingTV         object
StreamingMovies     object
Contract            object
PaperlessBilling    object
PaymentMethod       object
MonthlyCharges     float64
TotalCharges        object
Churn               object
dtype: object
```

Data Preprocessing: performed several steps mentioned below to ensure the data is ready for predictive modeling.

- **Feature Scaling:** Standardized categorical variables to simplify the categories and potentially improve model performance by reducing sparsity in the dataset.
- Reviewed and standardized categorical values in the 'MultipleLines' and columns related to 'Online Services' by merging 'No phone service' and 'No Internet Service' with 'No'. This ensures consistency and reduces redundancy in the data, making it more suitable for analysis and modeling.

- **Remove Customer IDs**: unique id which doesn't provide predictive power for the model.

- **Converting the Predictor Variable**: Converts the target variable 'Churn' from categorical ('Yes', 'No') to a binary numeric format (1, 0), which is necessary for binary classification models.

- **One-Hot Encoding of Categorical Variables**: df_dummies = pd.get_dummies(df2): This is the one-hot encoding step. Converted all categorical variables such as service types and demographic information in the dataset into dummy variables. Each categorical variable is now transformed into multiple binary variables (0 or 1) representing the different categories or levels of that variable.
- **Scaling Variables:** Utilized MinMaxScaler to scale features to a range of 0 to 1.

	tenure	MonthlyCharges	TotalCharges	Churn	gender_Female	gender_Male	SeniorCitizen_No	SeniorCitizen_Yes	Partner_No	Partner_Yes	...
0	1	29.85	29.85	0	1	0	1	0	0	1	...
1	34	56.95	1889.50	0	0	1	1	0	1	0	...
2	2	53.85	108.15	1	0	1	1	0	1	0	...
3	45	42.30	1840.75	0	0	1	1	0	1	0	...
4	2	70.70	151.65	1	1	0	1	0	1	0	...
5	8	99.65	820.50	1	1	0	1	0	1	0	...
6	22	89.10	1949.40	0	0	1	1	0	1	0	...
7	10	29.75	301.90	0	1	0	1	0	1	0	...
8	28	104.80	3046.05	1	1	0	1	0	0	1	...
9	62	56.15	3487.95	0	0	1	1	0	1	0	...

Exploratory Data Analysis:

Analysis of Numerical Features

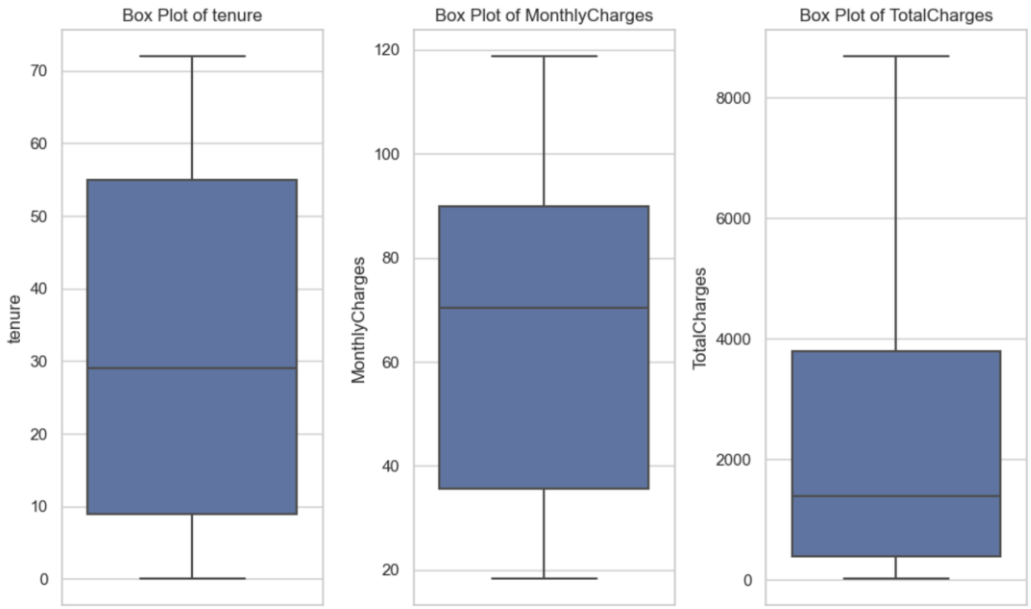
Box Plot Observations:

- **Tenure:** The box plot for tenure suggests that customers have a wide range of relationships with the company, from new to long-term customers. The distribution does not indicate the presence of outliers, implying a consistent customer retention pattern.
- **Monthly Charges:** Monthly charges are evenly spread, with the median indicating that most customers are charged a moderate amount on a monthly basis. The absence of outliers suggests that pricing strategies are consistent across different customer segments.

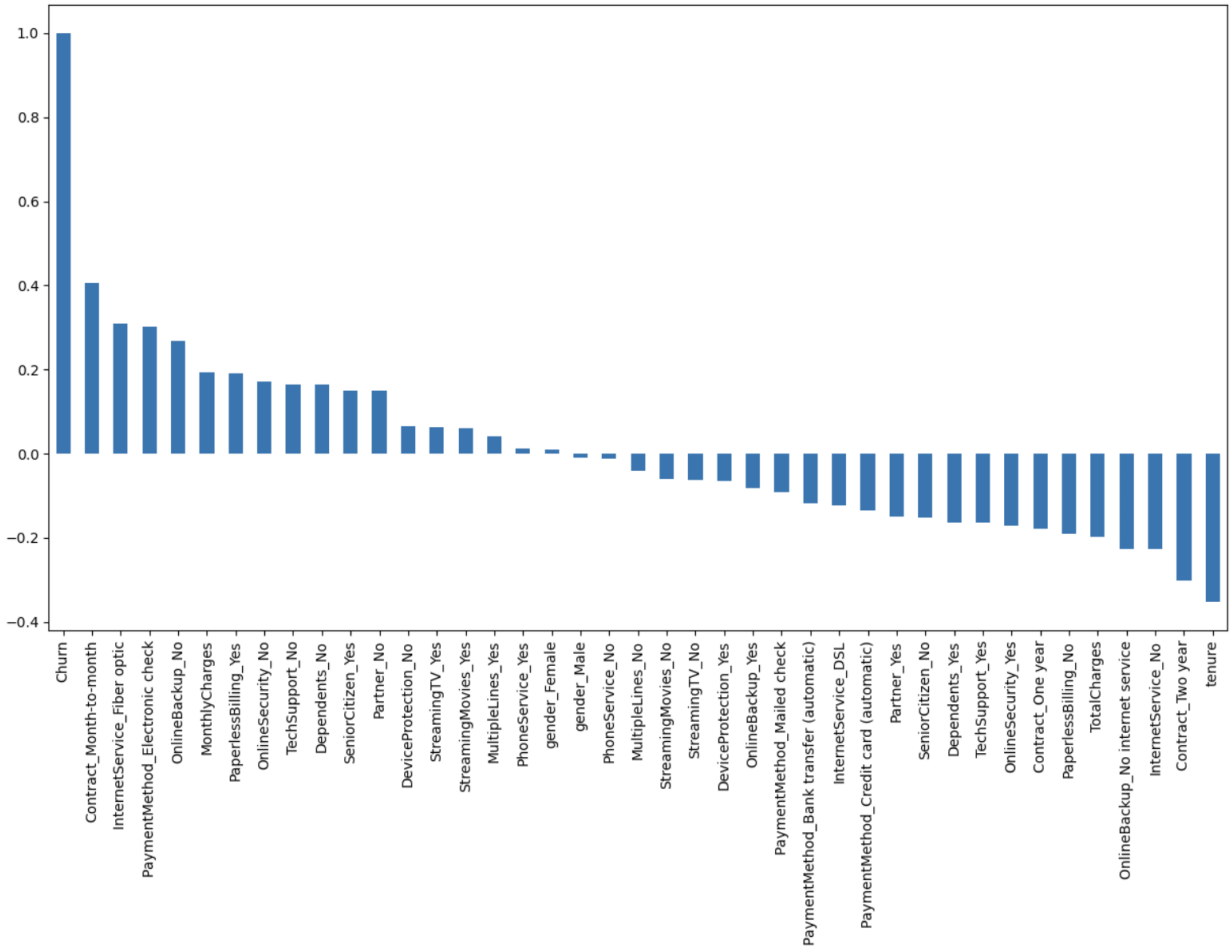
Insight on Total Charges:

- **Total Charges:** This box plot shows a more varied distribution, with a number of potential outliers indicated by points beyond the whiskers. These outliers represent customers with unusually high lifetime spendings.

The outliers could indicate customers who have either been with the company for a very long time or who have opted for premium services leading to higher lifetime payments.

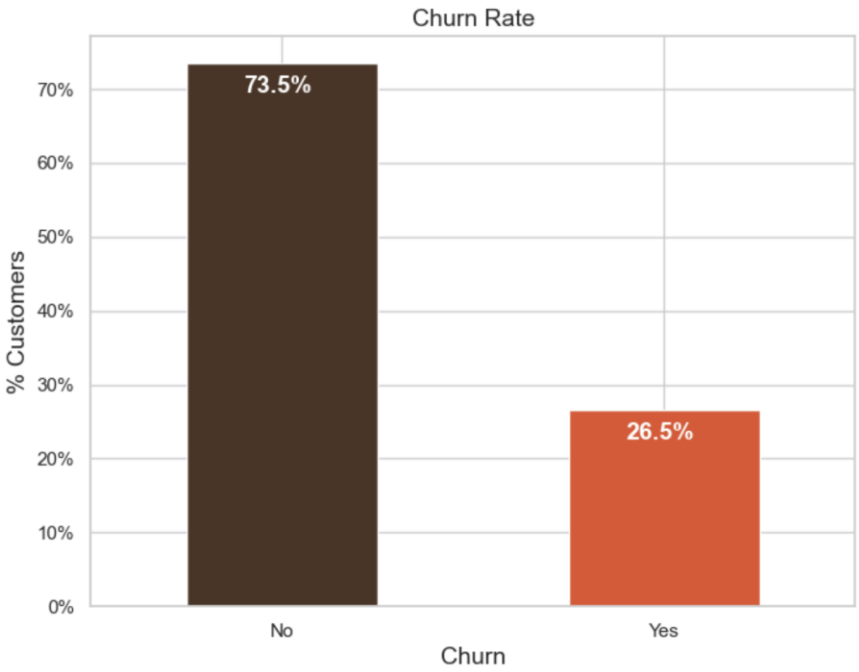


Correlation of "Churn" with other variables



- Month to month contracts, absence of online security and tech support seem to be positively correlated with churn. While, tenure, two year contracts seem to be negatively correlated with churn.
- Interestingly, services such as Online security, streaming TV, online backup, tech support, etc. without internet connection seem to be negatively related to churn.
- We will explore the patterns for the above correlations below before we delve into modelling and identifying the important variables.

Churn Rate & Distribution w.r.t Gender

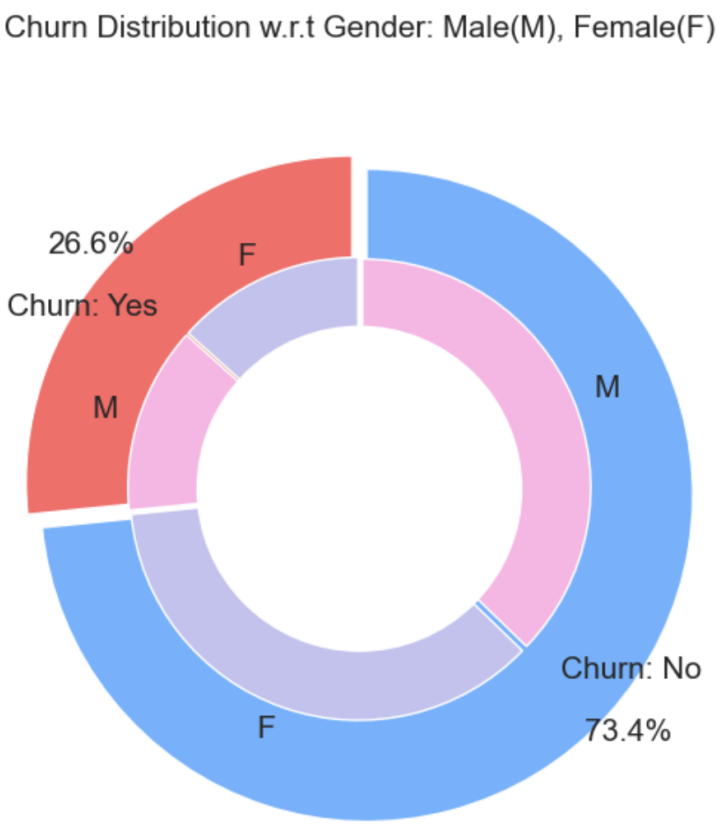


Churn Rate Analysis:

- **Overall Churn Rate:** We observe that 26.5% of our customer base has churned. This indicates that more than a quarter of the customers have discontinued service over the time period analyzed.

Gender Distribution in Churn:

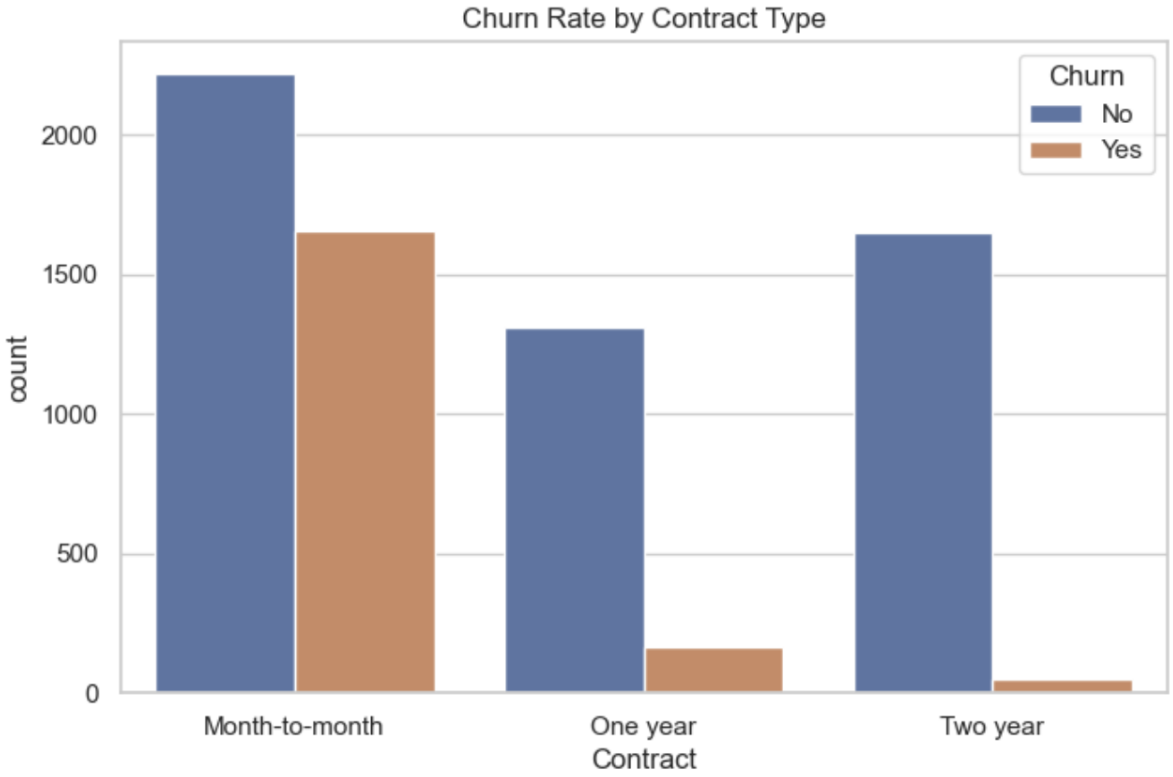
- **Churn by Gender:** The churn appears to be evenly distributed across genders. 26.6% of the churned customers are female, while males make up a nearly identical proportion.
- **Retention by Gender:** Similarly, gender distribution for retained customers (those who have not churned) is balanced, with both genders equally represented in the 73.5% of customers who continue to use our services.



Insights and Strategic Implications:

- **Churn Symmetry Across Genders:**
- The data suggests that gender does not play a significant role in customer churn for our services.
- This could indicate that factors leading to customer churn are independent of gender and are likely related to other aspects such as service quality, pricing, customer service, or product offerings.
- **Focus Areas for Retention Strategies:**
- Since gender does not significantly influence churn, retention efforts should be tailored based on other variables that show a stronger correlation with churn.
- It's essential to identify and address service aspects that contribute to customer dissatisfaction across all demographics.

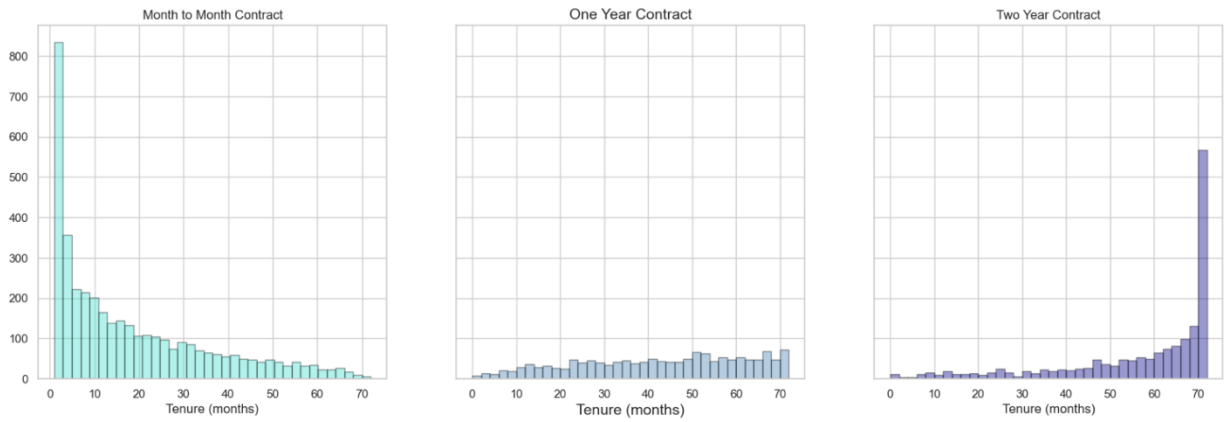
Churn Rate by Contract Type



Bar Chart Analysis:

- **Month-to-Month Contracts:** Customers with month-to-month contracts show the highest churn rate. This suggests a lower commitment level and potentially a response to short-term dissatisfaction or competitive offers.
- **One-Year Contracts:** Customers on one-year contracts show lower churn rates compared to month-to-month, indicating a moderate level of commitment.
- **Two-Year Contracts:** The churn rate for customers on two-year contracts is the lowest. This implies a high level of customer satisfaction and loyalty, as these customers are willing to commit to a longer term.

Contract Distribution



Month to Month Contract Histogram:

- The distribution is heavily right-skewed, indicating that most customers on a month-to-month contract have a short tenure.
- There is a significant drop after the first month, which suggests a high turnover rate; many customers leave after the first month.
- The numbers gradually decline as tenure increases, with very few customers remaining beyond 20 months.

One Year Contract Histogram:

- This distribution appears more uniform compared to the month-to-month contract, but still, there are fewer customers with higher tenure.
- There are peaks around 12 and 24 months, which could indicate that some customers renew their contract after it expires, but there is not a significant spike like in the two-year contract histogram.
- The data is somewhat sparse, with fewer customers in each tenure bin compared to the month-to-month contract.

Two Year Contract Histogram:

- This distribution shows a significant spike at 70 months, which is unusual. It suggests a large number of customers whose contracts have just reached the end of a two-year period, or possibly data that has been capped or recorded up to a certain tenure limit.
- Aside from the spike at the end, the rest of the histogram is quite sparse, indicating that fewer customers are at the intermediate tenure levels.

Insights Across All Contracts:

- The difference in distribution shapes indicates that customer retention varies significantly with the type of contract.
- Month to month contracts might be attracting customers who are not looking for long-term commitment and therefore have a high churn rate.
- One year contracts show more stable retention throughout the year, but not as much as two-year contracts.
- Two year contracts likely have the best retention, evidenced by the accumulation of customers at the end of the histogram, but this could also suggest that the data is right-censored (we do not see what happens beyond 70 months).

Feature Engineering

- Engineered a new feature called "HighTenureHighMonthly". The feature is set to 1 if both 'tenure' and 'MonthlyCharges' are above their respective medians for each observation, otherwise, it's set to 0.

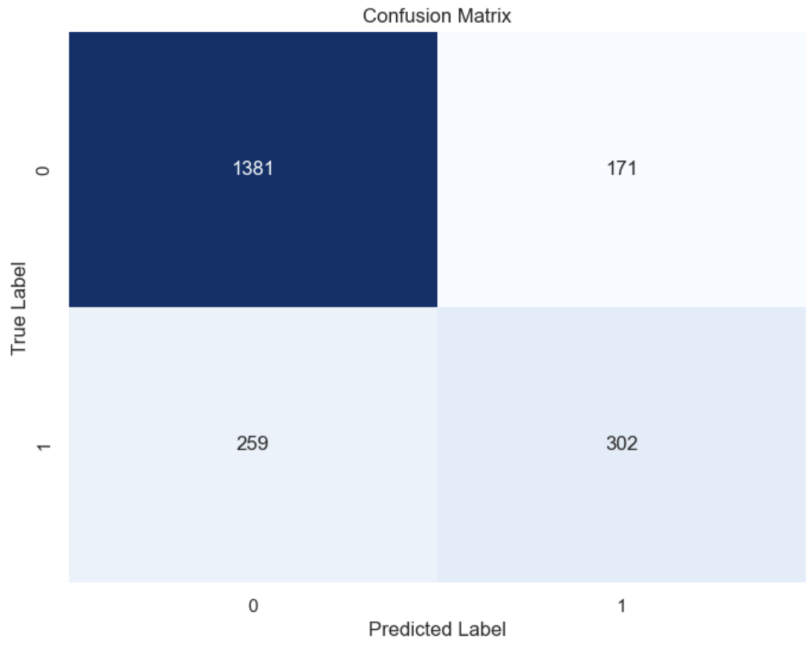
- Customers who have been with the company for a long time and pay more than the average monthly fee might be more loyal or have a higher satisfaction level, potentially making them less likely to churn. Conversely, they could also be more at risk if they find a better offer elsewhere, considering they might save more money.

- The model can then pick up on patterns that are specifically related to this segment of customers.

Splitting Data & Model Development

- For our analysis, we will utilize the data frame where dummy variables have been created for categorical features:
- To ensure uniformity and facilitate model convergence, we will scale all variables to a range of 0 to 1 using Min-Max Scaling.
- Next, we'll split the data into training and testing sets, allocating 30% of the data for testing while ensuring that the class distribution is preserved using stratified sampling.

Logistic Regression



Model Evaluation Metrics:

- **Data Split:** The dataset was divided into training (70%) and testing (30%) sets to evaluate the model's performance on unseen data.
- **Model Accuracy:** The model achieved an overall accuracy of 80%, which is the percentage of total correct predictions.

Confusion Matrix Insights:

- **True Negatives (TN):** 1381 predictions correctly identified customers who did not churn.
- **False Negatives (FN):** 259 predictions incorrectly identified customers who churned as non-churn.
- **True Positives (TP):** 302 predictions correctly identified customers who churned.
- **False Positives (FP):** 171 predictions incorrectly identified non-churn customers as churn.

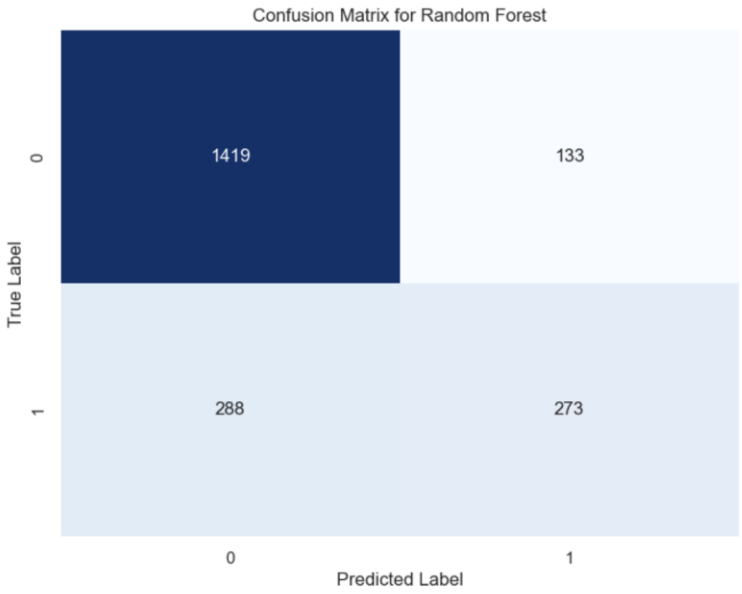
Classification Report:

- **Precision for Class 0 (No Churn):** The model has a precision of 0.84, meaning it is quite accurate when it predicts a customer will not churn.
- **Recall for Class 0 (No Churn):** The recall for non-churning customers is 0.89, indicating the model is good at identifying actual non-churn cases.
- **F1-Score for Class 0 (No Churn):** The F1-score for non-churning customers is 0.87, showing a balanced precision-recall trade-off for this class.
- **Precision for Class 1 (Churn):** The model has a lower precision of 0.64 for predicting churn, suggesting room for improvement in correctly identifying churn cases.
- **Recall for Class 1 (Churn):** The recall for churning customers is 0.54, indicating that the model misses almost half of the actual churn cases.
- **F1-Score for Class 1 (Churn):** The F1-score for churning customers is 0.58, reflecting the model's moderate performance for this class.

Overall Assessment and Next Steps:

- **Strengths:** The model is robust in identifying customers who will not churn, which is valuable for ensuring customer satisfaction efforts are not unnecessarily targeted at secure customers.

Random Forest



Evaluating Random Forest Model

Accuracy and Performance Metrics:

- **Accuracy Score:** The model has an accuracy score of approximately 80%, indicating a solid overall performance in predicting customer churn.

Confusion Matrix Interpretation:

- **True Negatives (TN):** 1419 cases where the model correctly predicted customers would not churn.
- **False Positives (FP):** 133 cases where the model incorrectly predicted churn for customers who did not actually churn.
- **False Negatives (FN):** 288 cases where the model failed to predict churn for customers who did churn.
- **True Positives (TP):** 273 cases where the model correctly predicted customers would churn.

Classification Report Summary:

- **Class 0 (No Churn):**
- **Precision:** With a precision of 0.83, the model is quite reliable in predicting non-churn cases.
- **Recall:** The recall of 0.91 for non-churn predictions indicates the model is highly capable of identifying actual non-churn instances.
- **F1-Score:** An F1-score of 0.87 for non-churn cases suggests a good balance between precision and recall for this class.
- **Class 1 (Churn):**
- **Precision:** A precision of 0.67 for churn predictions implies there's room to improve the model's ability to accurately predict actual churn cases.
- **Recall:** The recall of 0.49 for churn cases indicates the model is missing more than half of the actual churn cases.
- **F1-Score:** An F1-score of 0.56 for churn cases shows the model's limited effectiveness in predicting churn.

Critical Evaluation and Further Action:

- **Strengths:** The model performs well in predicting non-churning customers, which helps prevent unnecessary retention efforts for secure customers.

Cross Validation

- The cross-validation scores for Logistic Regression are fairly consistent across the folds, ranging from 0.789 to 0.813, with a mean accuracy of approximately 0.803.
- On the other hand, the Random Forest model has an out-of-bag (OOB) score of approximately 0.802, which is quite similar to the Logistic Regression mean accuracy obtained from cross-validation.

Based on the provided metrics, the Logistic Regression model seems to perform the best for customer churn prediction. While the Random Forest model has comparable accuracy to the Logistic Regression model, its lower recall suggests that it may miss more instances of actual churn compared to the Logistic Regression model.

Result & Interpretation

Logistic Regression:

- Having a 2 month contract reduces chances of churn. 2 month contract along with tenure have the most negative relation with Churn as predicted by logistic regressions
- Having DSL internet service also reduces the proability of Churn
- Lastly, total charges, monthly contracts, fibre optic internet services and seniority can lead to higher churn rates.

Random Forest:

- Monthly contract, tenure and total charges are the most important predictor variables to predict churn.
- The results from random forest are very similar to that of the logistic regression and in line to what we had expected from our EDA.