# "Customer Churn Prediction using Support Vector Machine (SVM)"

## Introduction

Customer churn prediction plays a crucial role in helping telecom companies identify customers who are likely to leave, enabling them to take proactive steps to retain these valuable customers. The goal of this analysis is to understand patterns of customer churn using a dataset that includes various customer attributes such as demographics, service usage, and billing information. Through exploratory data analysis (EDA), we aim to uncover insights into customer behavior and identify the key factors that influence churn. Following this, we apply a machine learning model, specifically a Support Vector Machine (SVM), to predict whether a customer will churn based on the extracted features. By combining EDA and predictive modeling, this project seeks to provide actionable insights that telecom companies can use to improve customer retention strategies.

## Analysis

The dataset consists of 7,043 observations and 38 variables, including both numerical and categorical data. Key variables include demographic details (like age, gender, marital status), service usage metrics (e.g., tenure, charges, internet type), and churn-related fields such as Customer.Status, Churn.Category, and Churn.Reason. The target variable for prediction is Customer.Status, which indicates whether a customer has churned or stayed. There are no duplicate Customer IDs in the dataset, indicating that each row represents a unique customer. The dataset contained blank values in several columns, particularly in Churn.Category and Churn.Reason for customers who did not churn. These missing values were appropriately filled. Removed irrelevant columns like Customer ID, location details (Latitude, Longitude, Zip Code, City) to focus on meaningful predictors.

### i. Descriptive statistics

```
> summary(numeric_vars)
      Age        Number.of.Dependents Number.of.Referrals Tenure.in.Months Avg.Monthly.Long.Distance.Charges Avg.Monthly.GB.Download
 Min.   :19.00   Min.   :0.0000       Min.   : 0.000      Min.   : 1.00    Min.   : 0.00                     Min.   : 0.00
 1st Qu.:32.00   1st Qu.:0.0000       1st Qu.: 0.000      1st Qu.: 9.00    1st Qu.: 9.21                     1st Qu.: 3.00
 Median :46.00   Median :0.0000       Median : 0.000      Median :29.00    Median :22.89                     Median :17.00
 Mean   :46.51   Mean   :0.4687       Mean   : 1.952      Mean   :32.39    Mean   :22.96                     Mean   :20.52
 3rd Qu.:60.00   3rd Qu.:0.0000       3rd Qu.: 3.000      3rd Qu.:55.00    3rd Qu.:36.40                     3rd Qu.:27.00
 Max.   :80.00   Max.   :9.0000       Max.   :11.000      Max.   :72.00    Max.   :49.99                     Max.   :85.00
 Monthly.Charge   Total.Charges    Total.Refunds    Total.Extra.Data.Charges Total.Long.Distance.Charges Total.Revenue
 Min.   :-10.00   Min.   :  18.8   Min.   : 0.000   Min.   :  0.000          Min.   :   0.00             Min.   :   21.36
 1st Qu.: 30.40   1st Qu.: 400.1   1st Qu.: 0.000   1st Qu.:  0.000          1st Qu.:  70.55             1st Qu.:  605.61
 Median : 70.05   Median :1394.5   Median : 0.000   Median :  0.000          Median : 401.44             Median : 2108.64
 Mean   : 63.60   Mean   :2280.4   Mean   : 1.962   Mean   :  6.861          Mean   : 749.10             Mean   : 3034.38
 3rd Qu.: 89.75   3rd Qu.:3786.6   3rd Qu.: 0.000   3rd Qu.:  0.000          3rd Qu.:1191.10             3rd Qu.: 4801.15
 Max.   :118.75   Max.   :8684.8   Max.   :49.790   Max.   :150.000          Max.   :3564.72             Max.   :11979.34
```

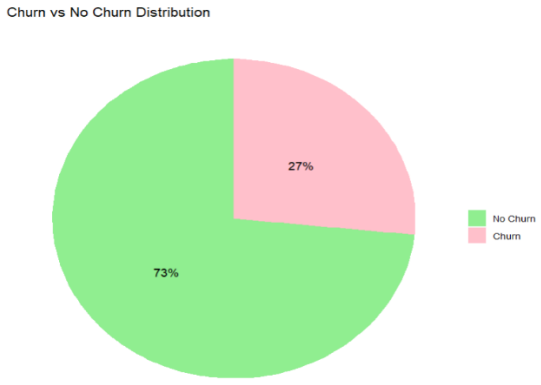
The dataset represents a diverse customer base with a wide range of usage and billing behaviors. Customers range in age from 19 to 80, with most having no dependents or referrals, suggesting a

majority are individual users. The average tenure is about 32 months, but spans from just 1 to 72 months, indicating both new and long-term users are present. Monthly charges range broadly from -$10 (likely an error) to $118.75, with a mean of $63.60. While the average Total Charges and Total Revenue are high ($2,280 and $3,034 respectively), the wide range and high maximums (up to $11,979) point to substantial variability in customer value. Most customers incur no refunds or extra data charges, suggesting these are edge cases but potentially meaningful when modeling churn or profitability.

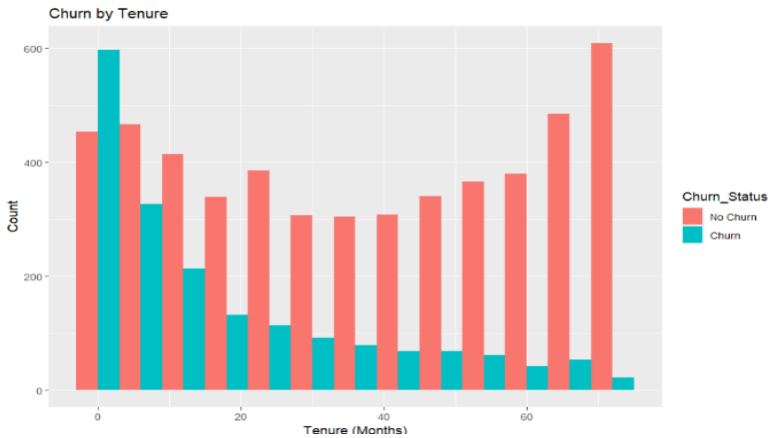
## ii. Visualization

- **Class Balance**

  This pie chart illustrates that, a substantial majority (73%) of customers did not churn, while a smaller portion (27%) did. This noticeable imbalance between the two classes is an important characteristic of the data to consider when building and evaluating predictive models for churn.



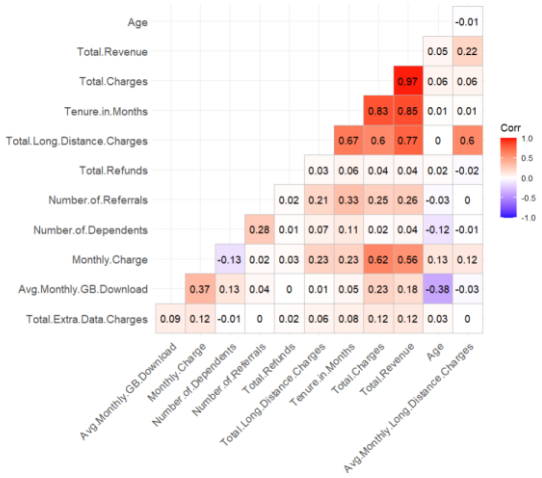

- **Churn by Tenure**




The bar chart "Churn by Tenure" reveals an interesting trend: customers with very short tenures (0-10 months) show a high propensity to churn, as do customers around the 70-month mark. In contrast, customers who have stayed for a mid-range duration (approximately 20-60 months) exhibit lower churn rates. This pattern indicates that customer tenure is a significant predictor of

churn, with both early dissatisfaction and potentially end-of-service factors contributing to higher churn at the extremes of tenure.

- **Correlation**



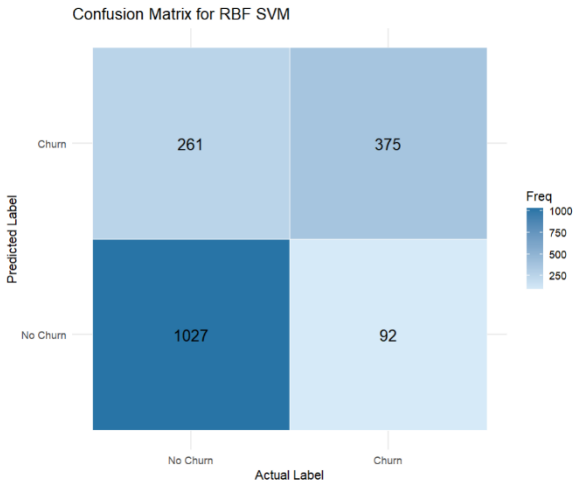

Total Revenue is strongly correlated with Total Charges (0.97), Tenure in Months (0.85), and Total Long-Distance Charges (0.77), indicating these are key drivers of revenue. Most other variables show weak or negligible correlations, except for some moderate ones like Monthly Charge with Total Charges (0.62) and Average Monthly GB Download with Monthly Charge (0.37).

### iii. SVM Model with Radial Basis Function (RBF) Kernel

In this analysis, a Support Vector Machine (SVM) model with a Radial Basis Function (RBF) kernel is used to predict customer churn. The SVM model is a powerful supervised machine learning algorithm that is particularly effective for classification tasks, such as predicting whether a customer will churn or not, based on multiple attributes.

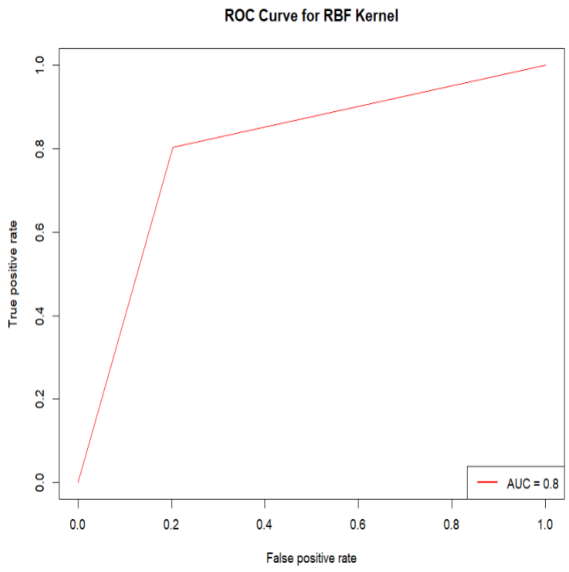
### iv. Confusion Matrix

```
> # Confusion Matrix for the RBF SVM
> confusionMatrix(predictions_rbf, test_set$Churn_Status)
Confusion Matrix and Statistics

          Reference
Prediction No Churn Churn
  No Churn     1027    92
  Churn         261   375

               Accuracy : 0.7989
                 95% CI : (0.7793, 0.8174)
    No Information Rate : 0.7339
    P-Value [Acc > NIR] : 1.406e-10

                  Kappa : 0.5383

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7974
            Specificity : 0.8030
         Pos Pred Value : 0.9178
         Neg Pred Value : 0.5896
             Prevalence : 0.7339
         Detection Rate : 0.5852
   Detection Prevalence : 0.6376
      Balanced Accuracy : 0.8002

       'Positive' Class : No Churn
```



The confusion matrix and associated metrics indicate that the model performs well, with an accuracy of 79.89%, suggesting it correctly predicts the outcome in most cases. The model is particularly good at identifying "No Churn" cases, with a high positive predictive value (91.78%) and specificity (80.30%). However, there is room for improvement in predicting "Churn" cases, as reflected in the relatively lower negative predictive value (58.96%) and the occurrence of false negatives (261 instances of "No Churn" predicted as "Churn"). The balanced accuracy of 80.02% further reinforces the model's overall good performance, but enhancing its ability to correctly classify "Churn" could make it even more effective. The moderate Kappa value (0.5383) shows the model's agreement with actual outcomes is better than random guessing.

v. **AUC Score**:




An AUC of 0.8 means the model has strong discriminatory power, correctly distinguishing between the "No Churn" and "Churn" classes most of the time. This value indicates that the model is quite reliable, as it suggests that there's an 80% chance that the model will correctly rank a randomly selected "No Churn" instance higher than a randomly selected "Churn" instance. The fact that the curve rises steeply and gets close to the top-left corner is crucial because it reflects high sensitivity (True Positive Rate) and low false positive rate, which is ideal for models in classification tasks like this.

## Conclusion

In this project, we analyzed customer churn in a telecom company using machine learning techniques. Through Exploratory Data Analysis (EDA), we identified key factors influencing churn, such as tenure and total charges. The SVM model with an RBF kernel achieved an accuracy of 79.89%, demonstrating good performance in predicting non-churning customers. However, the model struggled to predict churned customers effectively, showing better specificity but lower sensitivity. Future work should focus on balancing the data using techniques like SMOTE, adjusting class weights, and exploring alternative models, such as ensemble methods or neural networks, to improve churn prediction accuracy. Additionally, hyperparameter tuning, feature engineering, and model interpretability could further enhance model performance and provide deeper insights into customer behavior.