Telecom Churn Analysis

**Adhoksh Milind Sonawane Sakshi Sanjay Dudhal Kaustubh Arun Pasalkar San Diego State University**

**MIS 720: Electronic Business and Big Data Infrastructures Dr. Xialu Liu**

**6th May 2024**

**Table of Contents**

1. **Introduction**
   * Overview of Customer Retention Importance 3
   * Objective of the Study 3
2. **Data Preparation**
   * Summary of Dataset 4
   * Data Cleaning: Handling Missing Values 6
   * Data Transformation 7
3. **Exploratory Data Analysis (EDA)**
   * Descriptive Statistics 7
   * Visualization Techniques 7
4. **Model Building and Selection**
   * Feature Selection 7
   * Data Splitting 7
   * Classification Approach Using SVM 10
     + Linear SVM 10
     + Polynomial SVM 11
     + Radial SVM 12
5. **Model Planning** 
   * Implementation Strategy 13
   * Kernel Type Comparison 13
6. **Strategies to Reduce Churn** 
   * Customer Service Improvement 14
   * Pricing Strategy Adjustment 14
   * Enhancing Customer Onboarding and Retention 14
7. **Conclusion** 15
8. **References** 15

**Telecom Churn Analysis**

Customer retention has emerged as a critical component for companies looking to sustain profitable growth in today's very competitive market. The proliferation of subscription-based services in a variety of sectors has made it critical to comprehend and forecast consumer behavior, especially churn. Businesses may use targeted retention methods to improve customer satisfaction and loyalty by being able to predict when a client is likely to leave.

Using the Churn Dataset from Orange Telecom, we explore the field of predictive modeling in this research. This dataset includes customer activity data that has been painstakingly cleansed and enhanced with a critical churn label that indicates if a customer has cancelled their subscription. Our goal is to create strong prediction models that can anticipate churn tendencies in Orange Telecom's customer base by utilizing machine learning.

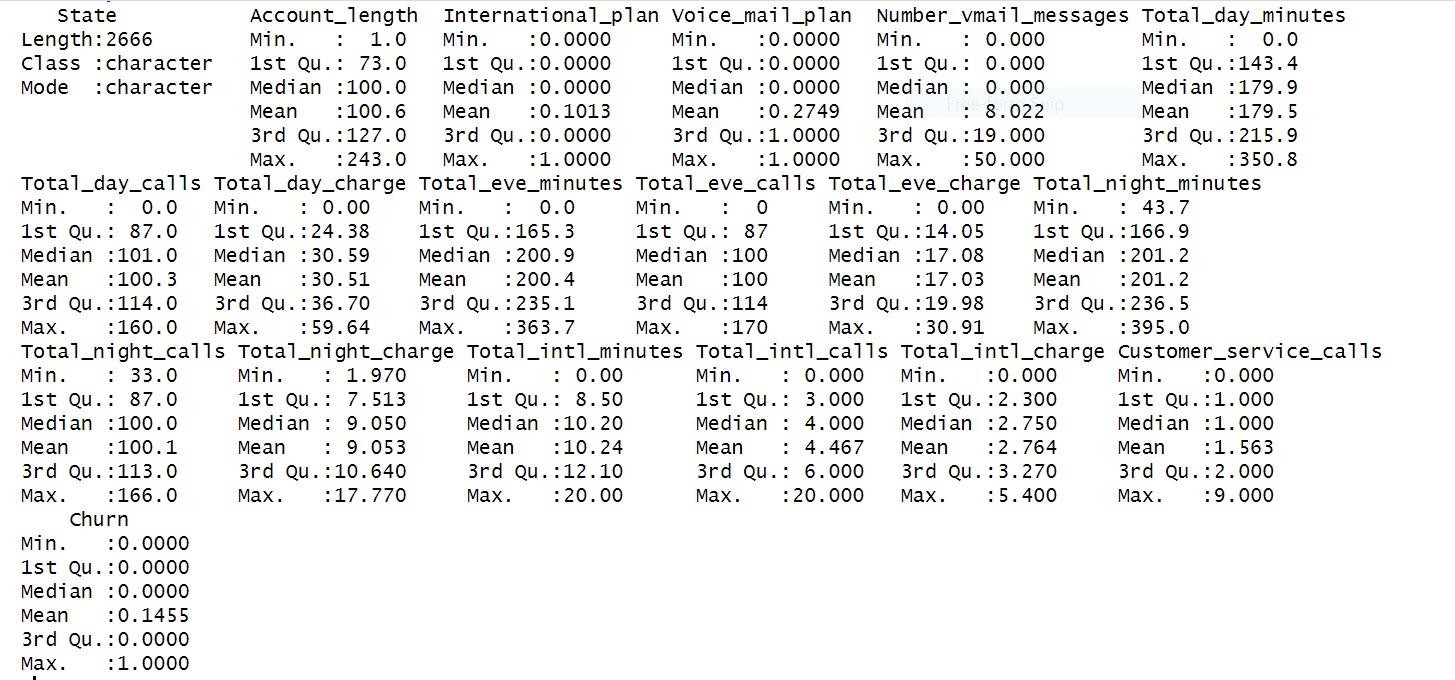
This project offers a chance to explore and learn about customer churn prediction and is a useful example of predictive analytics in action. Through a thorough examination and analysis of Orange Telecom's Churn Dataset, our goal is to provide valuable information that may guide strategic choices and stimulate creative approaches to customer retention.

This dataset contains various attributes related to customer service usage and preferences. State refers to the U.S. state where the customer resides. Account Length measures how long, in months, the customer has been with the service. International Plan and Voice Mail Plan indicate whether the customer has international calling or voicemail services, respectively. The Number of Voicemail Messages shows how many voicemails are stored. Total Day, Evening, and Night Minutes and Calls record the duration and number of calls made during different times of the day, while Total Day, Evening, and Night Charges show the associated costs. Total International Minutes, Calls, and Charges track the customer's international calling usage and expenses.

Customer Service Calls count the number of times customer service was contacted. Finally, Churn shows whether the customer has discontinued the service. These variables collectively provide a comprehensive overview of customer activity and service usage.

# Data Preparation

Summary of Dataset



This document contains the telecom dataset's summary statistics, which include 2,666 customer records. The dataset contains a variety of user information, including consumption measurements at different times of the day, service plans, state, and account term. For every property, important metrics are outlined, including the lowest, first quartile, median, mean, third quartile, and maximum values. The attributes such as `Total\_day\_minutes`, `Total\_day\_calls`,

`Total\_day\_charge`, `Total\_eve\_minutes`, `Total\_eve\_calls`, `Total\_eve\_charge`,

`Total\_night\_minutes`, `Total\_night\_calls`, `Total\_night\_charge`, `Total\_intl\_minutes`,

`Total\_intl\_calls`, `Total\_intl\_charge`, and `Customer\_service\_calls` are especially important as they provide information on client usage trends and interactions with services. The large range of daily usage (up to 350.8 minutes) and the number of customer care calls (up to 9) that indicate different degrees of consumer inquiries or complaints are noteworthy facts.

# Table 1

*Central tendencies and dispersion of customer activity*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
| Total\_day\_minutes | 0.0 | 143.7 | 179.5 | 179.5 | 215.9 | 350.8 |
| Total\_day\_calls | 0 | 87 | 101 | 100.3 | 114 | 160 |
| Total\_day\_charge | 0.0 | 24.43 | 30.51 | 30.51 | 36.70 | 59.64 |
| Customer\_service\_  calls | 0 | 1.0 | 1.0 | 1.56 | 2.0 | 9.0 |

The table above shows a subset of these statistics that can be crucial for predictive analytics in customer attrition and service improvement. It also shows normal customer behavior and service consumption.

# Correlation Matrix

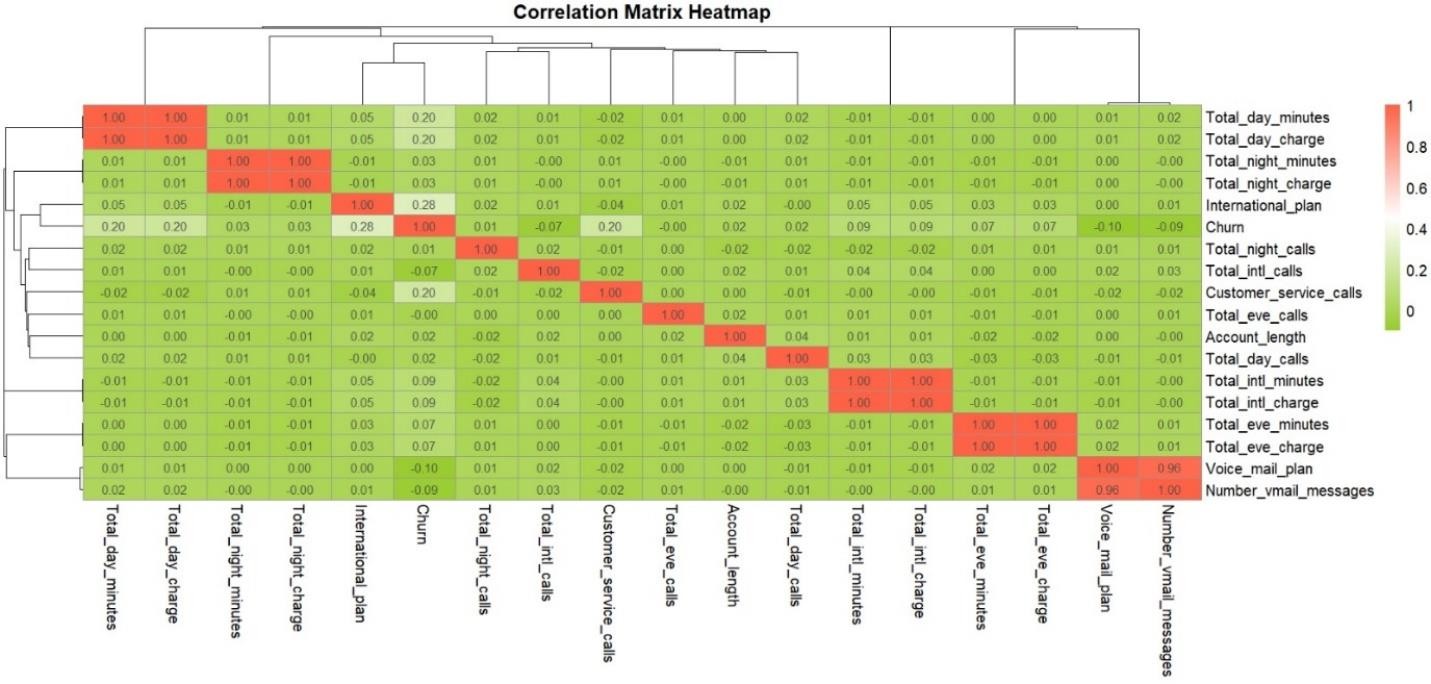


Figure 1

We can examine the links between various variables in the correlation matrix. The varied correlation coefficients indicate different amounts of impact on consumer choices. Based on the greatest positive values across all the variables, this heatmap identifies probable churn predictors.

As a result, we can track **Total day minutes, Total day charge, international plans, and Customer Service calls** which help us determine the best areas to target with our retention campaigns. Additionally, there are some negative correlation values for the voice mail plan, the quantity of voicemail messages, and the total number of foreign calls. The inverse relationship between churn and each of these indicators is shown by the negative number. For instance, if the overall number of foreign calls rises, the Churn correspondingly falls. However, as the values are not too low.

# Data Collection and Scope

- Source Identification: <https://www.kaggle.com/datasets/mnassrib/telecom-churn-datasets/data> The dataset used for the churn analysis encompasses a range of customer attributes, including service usage details and various plan options. These attributes provide a critical foundation for analyzing the factors contributing to customer churn. By examining variables such as the frequency of service usage and the types of plan options selected by customers, we can identify patterns and trends that indicate potential dissatisfaction or contentment among the customer base. This analysis helps in pinpointing the key drivers of customer churn and allows for the development of targeted strategies to improve customer retention.

1. **Data Cleaning -** Handling Missing Values

In the data cleaning phase, all missing values within the dataset were removed to ensure the integrity and accuracy of the subsequent analyses. This step is crucial as missing data can lead to biased or incorrect conclusions if not addressed appropriately. Despite the presence of three distinct values, preliminary analysis indicated that the '**Area Code**' did not significantly influence churn outcomes, making it an extraneous variable in the context of churn prediction. By removing '**Area Code**' from the analysis, we could focus on more impactful variables directly linked to customer behavior and service satisfaction, thereby simplifying the model and potentially improving its predictive performance.

# Data Transformation

To streamline the analysis process, categorical variables such as **International\_plan** and **Voice\_mail\_plan** were transformed from string representations ("No", "Yes") into binary format (0, 1). This encoding simplifies the input into analytical models, making it easier to handle and interpret. Converting these variables into a binary format eliminates any ambiguity associated with string processing and enhances the efficiency of the modeling process, as numerical data is more straightforward for algorithms to process and analyze. This transformation is essential for applying machine learning techniques that require numerical input.

# Exploratory Data Analysis (EDA)

**Descriptive Statistics**: The report includes plotting histograms for quantitative variables to visualize data distribution and plotting pie charts for qualitative variables to observe proportions.

**Visualizations**: Utilized `ggplot2` for generating histograms and `pie ()` function for pie charts. Correlation among numeric variables is explored through heatmaps using `heatmap`, highlighting potential multicollinearity or relationships.

Before diving into churn prediction, the initial step involves preparing the data to ensure high quality and reliability of the results. This process includes:

**Feature Selection**: Selecting relevant features that contribute to understanding customer churn, such as call charges, plan types, and service interactions.

**Data Splitting**: Dividing the data into training and testing sets with an 80-20 split to evaluate the model's performance on unseen data.

# Software and Tools Used

Data Handling Tools : The R programming language is used, leveraging libraries such as `dplyr` for data manipulation, `ggplot2` for data visualization, `heatmap` for heatmaps, `corrplot` for plotting correlation matrices, and `readxl` for reading Excel files.

# Quantitative Variables Depicted with Histograms and Box plots:

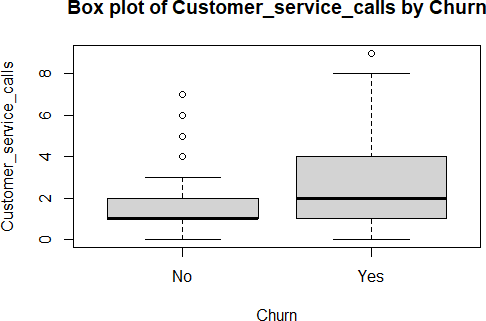
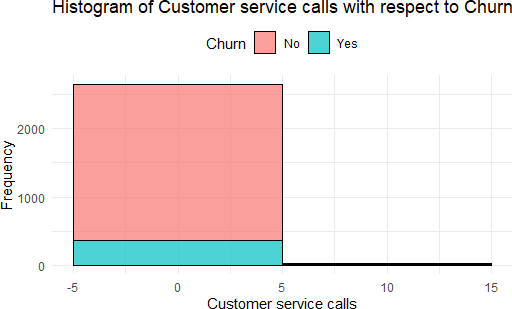


Figure 2

This analysis reveals a notable pattern in customer behavior related to churn. Churned customers are characterized by a higher volume of calls, as indicated by a wider distribution in the box plot. This suggests that these customers may have been experiencing issues or dissatisfaction leading them to reach out more frequently. Conversely, non-churned customers exhibit fewer calls, reflected in a tighter distribution in their box plot. This trend underscores that a higher number of calls might be a potential indicator of underlying problems that could eventually contribute to customer churn.

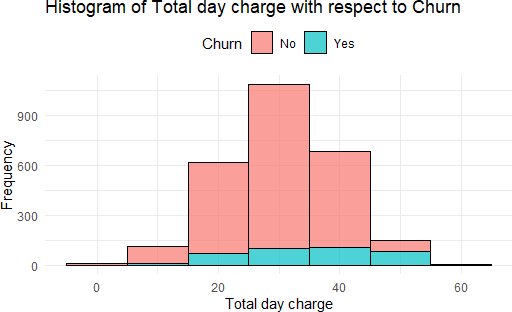
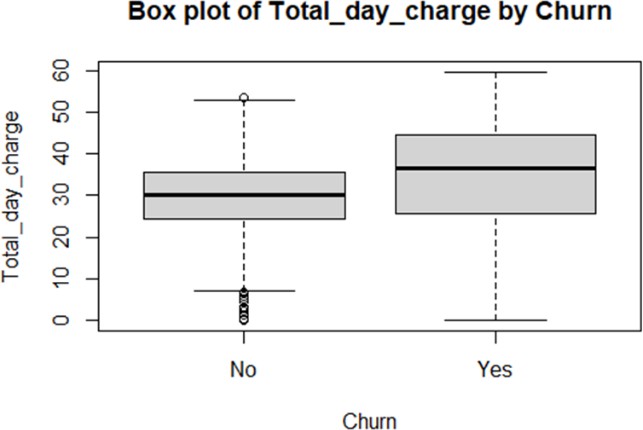


Figure 3

Comparing total day charges for non-churned and churned customers, the given visualizations show that day charges for non-churned customers are typically lower and more consistent.

Churned clients, on the other hand, typically have more diversified and higher fees. Pricing may

play a big role in client retention efforts, since the histogram and box plot collectively imply that higher day costs are linked to a larger chance of customer churn.

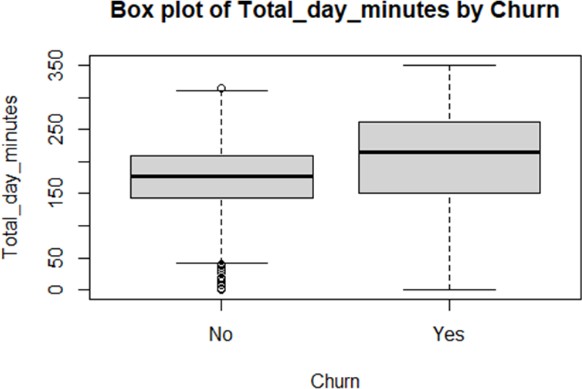
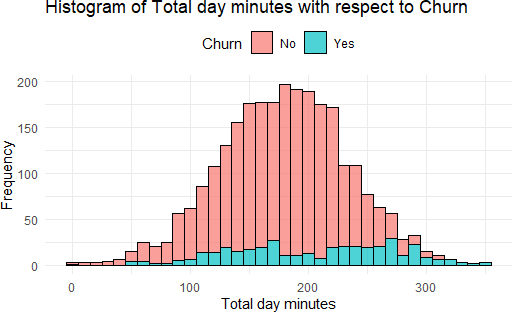


Figure 4

Both churned and non-churned clients often utilize a comparable amount of total daily minutes, centered around 200 minutes, as seen by the histograms and box plots. Customers that have not experienced churn exhibit a highly concentrated peak consumption, whereas those who have seen churn exhibit a little decrease in instances and a moderate skew toward lower minute usage.

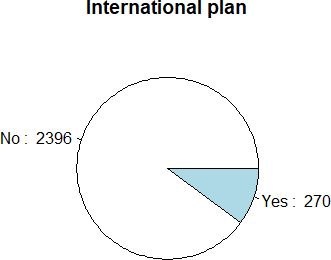
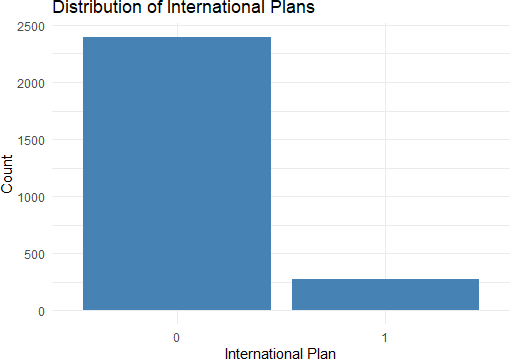


Figure 5

The pie chart and bar plot show how clients are divided up among international plans. The bar chart shows a discrepancy between the number of consumers who subscribe to an international plan (270) and most customers (2396), as seen by the wide bar at "0" and the considerably

smaller bar at "1". This information is further supported by the pie chart, which shows a large portion of consumers ("No" - 2396 customers) who do not have an overseas plan and a smaller portion ("Yes" - 270 customers) who do. These visual aids make it very evident that fewer customers choose the international package.

**Implementation Strategy**

The research makes use of comprehensive customer data and the Support Vector Machine (SVM) technique to forecast client attrition. Customer interactions, usage statistics, and account details including call costs and plan kinds are all included in this data. By analyzing these inputs, the SVM algorithm learns how to divide consumers into two categories: "churn" and "no churn." This is accomplished by finding the ideal hyperplane that optimizes the margin between each class's nearest data points while also dividing the two classes.

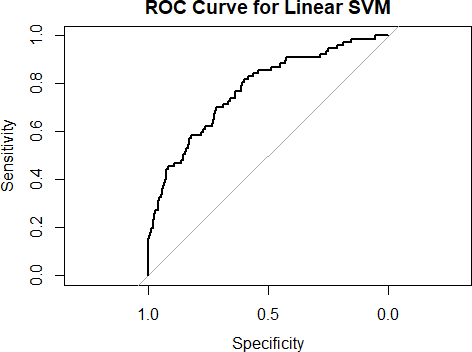
# Classification Approach Using SVM

Utilizing Support Vector Machine (SVM) models, customers are categorized as either "churn" or "no churn." The application of various SVM models is as follows:

1. Linear SVM
2. Polynomial SVM
3. Radial SVM

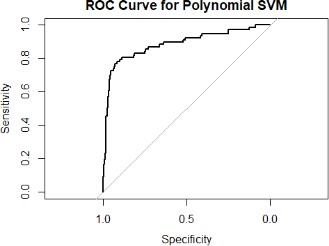
**Linear SVM:** Using a linear kernel to comprehend simple correlations in the data, this model acts as the standard. It can be restricted in its ability to capture more intricate patterns, but it is useful for obtaining first ideas. The ROC curve for the Linear SVM shows a decent performance with the curve staying significantly above the diagonal line, which represents a random guess.

However, the progression towards the top left corner is not very steep, suggesting moderate



sensitivity and specificity across various threshold levels.This indicates that the linear kernel provides an acceptable level of discrimination between the two classes. The Linear SVM kernel achieves an accuracy of 86.0% and an AUC of 0.8. The misclassification rate stands at 14.0%. These figures indicate a decent performance, particularly suitable for datasets with linearly separable features. The AUC value shows good but not excellent discriminatory ability.

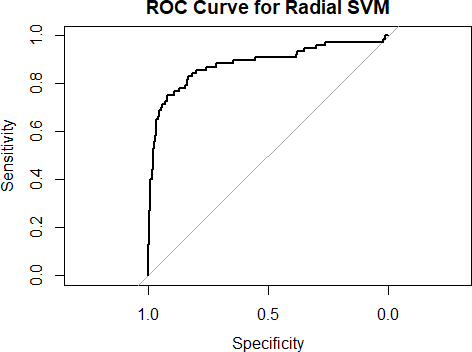
**Polynomial SVM**: This model interprets non-linear data patterns more accurately than the linear model thanks to the addition of a polynomial kernel. It adapts more skillfully to the subtleties of consumer interactions and behaviors.



The Polynomial SVM displays a better performance compared to the Linear SVM. The curve is closer to the top left corner of the plot, which indicates higher true positive rates (sensitivity) for corresponding false positive rates (1 - specificity). This suggests that the polynomial kernel, which considers interaction terms between features up to a certain degree, captures more

complex patterns in the data, thus improving classification accuracy. The Polynomial SVM also registers an accuracy of 86.0%, but with a higher AUC of 0.9. The misclassification rate remains the same as the Linear SVM at 14.0%. The higher AUC value suggests that the Polynomial kernel is better at discriminating between the classes, likely due to its ability to handle non-linear data interactions better than the Linear kernel.

**Radial SVM (RBF)**: Complex and multi-dimensional data patterns are best handled by this form of SVM, which makes use of a radial basis function (RBF) kernel. It excels in working with complex and diverse data structures. The ROC curve for the Radial SVM is the closest to the top left corner among the three.This demonstrates a high sensitivity, meaning it correctly identifies a high proportion of actual positives. The radial kernel, which can handle the case when the relationship between class labels and attributes is nonlinear, shows the highest overall effectiveness in distinguishing between the classes.



The Radial SVM shows the best performance among the three with an accuracy of 90.0% and an AUC of 0.9. It also has the lowest misclassification rate at 10.0%. This indicates that the Radial kernel is particularly effective in complex datasets where the relationship between features and classes is highly non-linear. The high AUC value further highlights its superior capability to differentiate between class labels effectively.

|  |  |  |  |
| --- | --- | --- | --- |
| **Kernel Type** | **Accuracy** | **Misclassification Rate** | **AUC** |
| Linear SVM | 86.0% | 14.0% | 0.8 |
| Polynomial SVM | 86.0% | 14.0% | 0.9 |
| Radial SVM | 90.0% | 10.0% | 0.9 |

# Key Takeaway:

The Radial SVM model outperforms the linear and polynomial counterparts, showcasing its superior predictive capabilities in identifying customers at risk of churning. This model's effectiveness underscores the potential of machine learning to empower proactive strategies in managing customer retention, making it an invaluable tool in the battle against customer churn. By leveraging the strengths of Radial SVM, businesses can accurately pinpoint and address the factors leading to customer attrition, thus enhancing customer satisfaction and loyalty.

The ROC curves for three different Support Vector Machine (SVM) kernels—linear, polynomial, and radial (also known as Gaussian or RBF)—are depicted in the provided graphs. These curves are instrumental in evaluating the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Here's a brief analysis of each:

In conclusion, while all three kernels demonstrate good performance metrics, the **Radial SVM** stands out as the most effective for the dataset in question, offering the highest accuracy and the ability to handle complex patterns more efficiently. The choice between these kernels should consider the specific characteristics of the dataset and the required precision of the model

**Strategies to reduce churn:**

To reduce customer churn, businesses can implement several strategies based on key factors that influence why customers leave:

Customer Service Calls: Many calls to customer service often means customers are unhappy. Improving the quality of customer support and quickly resolving issues can help reduce churn. Training customer service representatives to handle calls more effectively and efficiently might prevent customers from feeling frustrated and leaving.

Call Charges: High charges for day, evening, night, and international calls can lead customers to consider other options. Reviewing pricing strategies and perhaps lowering costs or offering more tailored pricing plans can help retain customers. Promotions or loyalty discounts can also make customers feel valued and less likely to switch to a competitor.

Account Length: Customers who have been with the company for a shorter period are often more likely to leave. To address this, companies can focus on the onboarding process for new customers to ensure they feel welcome and fully understand the benefits of their services.

Providing exceptional service from the start and offering incentives to stay longer can also increase customer loyalty.

# References

1. Mnassri, B. (2019). Churn analysis of a telecom company. Retrieved from [Kaggle](https://www.kaggle.com/datasets/mnassrib/telecom-churn-datasets/data)
2. Gulati, A. P. (2022, March 15). Churn analysis of a telecom company. Analytics Vidhya. Retrieved from [https://www.analyticsvidhya.com/blog/2022/01/churn-analysis-of-a-](https://www.analyticsvidhya.com/blog/2022/01/churn-analysis-of-a-telecom-company/) [telecom-company/](https://www.analyticsvidhya.com/blog/2022/01/churn-analysis-of-a-telecom-company/)
3. Nhu Y., N., Van Ly, T., & Vu Truong Son, D. (2022, May 24). Churn prediction in the telecommunication industry using kernel Support Vector Machines. PLOS ONE. <https://doi.org/10.1371/journal.pone.0267935>
4. Nurtriana, A., Rachmawati, D. D., Artiyasa, M., & Sidiq, D. S. Z. (2024). Churn prediction analysis of telecom customers using svm, random forest, and logistic regression models using orange data mining tools. E3S Web of Conferences, 2024(31). Retrieved from [https://www.e3s-](https://www.e3s-conferences.org/articles/e3sconf/pdf/2024/31/e3sconf_iccsei2023_02012.pdf) [conferences.org/articles/e3sconf/pdf/2024/31/e3sconf\_iccsei2023\_02012.pdf](https://www.e3s-conferences.org/articles/e3sconf/pdf/2024/31/e3sconf_iccsei2023_02012.pdf)
5. Rodan, A., Faris, H., Alsakran, J., & Al-Kadi, O. (2014). A support vector machine approach for churn prediction in telecom industry. Information—An International Interdisciplinary Journal, 17(8). Retrieved from [https://www.researchgate.net/publication/264534919\_A\_Support\_Vector\_Machine\_Appr](https://www.researchgate.net/publication/264534919_A_Support_Vector_Machine_Approach_for_Churn_Prediction_in_Telecom_Industry) [oach\_for\_Churn\_Prediction\_in\_Telecom\_Industry](https://www.researchgate.net/publication/264534919_A_Support_Vector_Machine_Approach_for_Churn_Prediction_in_Telecom_Industry)
6. Third Eye Data. (n.d.). Customer churn prediction with SVM using Scikit-Learn. Retrieved from [https://thirdeyedata.ai/customer-churn-prediction-with-svm-using-scikit-](https://thirdeyedata.ai/customer-churn-prediction-with-svm-using-scikit-learn/) [learn/](https://thirdeyedata.ai/customer-churn-prediction-with-svm-using-scikit-learn/)
7. Brandusoiu, I. (n.d.). Churn prediction in the telecommunications sector using support vector machines. Retrieved from Academia.edu: [https://www.academia.edu/57533676/Churn\_Prediction\_in\_the\_Telecommunications\_Se](https://www.academia.edu/57533676/Churn_Prediction_in_the_Telecommunications_Sector_Using_Support_Vector_Machines) [ctor\_Using\_Support\_Vector\_Machines](https://www.academia.edu/57533676/Churn_Prediction_in_the_Telecommunications_Sector_Using_Support_Vector_Machines)
8. Frohböse, F. (n.d.). Machine learning case study: Telco customer churn prediction. Towards Data Science. Retrieved from [https://towardsdatascience.com/machine-learning-](https://towardsdatascience.com/machine-learning-case-study-telco-customer-churn-prediction-bc4be03c9e1d) [case-study-telco-customer-churn-prediction-bc4be03c9e1d](https://towardsdatascience.com/machine-learning-case-study-telco-customer-churn-prediction-bc4be03c9e1d)