**Project overview**[**¶**](http://localhost:8888/notebooks/student.ipynb#Project-overview)

Churn refers to the number of customers who discontinue using a company’s services. In this context, it specifically pertains to customers who stop using SyriaTel’s telecommunications services and switch to other providers. Churn refers to the number of customers who discontinue using a company’s services. In this context, it specifically pertains to customers who stop using SyriaTel’s telecommunications services and switch to other providers.

This project aims to develop an advanced predictive model to identify customers at high risk of imminent churn from SyriaTel, a leading telecommunications provider in Syria. Structured as a binary classification problem, the model will assess the likelihood of a customer discontinuing their service. By leveraging historical customer behavior data to generate actionable insights, the solution seeks to enable SyriaTel to proactively address retention challenges, enhance customer loyalty, and implement effective, data-driven strategies to mitigate churn.

**Business Understanding**

Develop a predictive model to identify at-risk SyriaTel customers, enabling proactive strategies to minimize churn and enhance loyalty. By leveraging historical data, the model will uncover actionable insights to address churn drivers, reduce revenue loss, and optimize retention efforts.

Tentative expected outcome is increased retention, efficient resource allocation, and improved customer experience. This initiative positions SyriaTel to stay competitive with data-driven retention strategies.

**Problem Statement**  
 SyriaTel seeks to predict customer churn to reduce revenue loss, improve retention, and optimize resource allocation, ensuring competitiveness in the telecom market

**Data Understanding**

The dataset analyzed originates from Kaggle and is available for direct download in CSV format. It comprises 3,333 entries and 21 columns representing customer data. The data types of the columns are as follows:

* 8 columns of type int64
* 8 columns of type float64
* 4 columns of type object
* 1 column of type bool

Feature/ Column Descriptions

* State: The U.S. state where the customer resides.
* Account Length: Duration (in days) of the customer’s account with the company.
* Area Code: The customer’s area code.
* Phone Number: The customer’s phone number.
* International Plan: Indicates whether the customer has subscribed to the international calling plan (True/False).
* Voice Mail Plan: Indicates whether the customer has subscribed to the voicemail plan (True/False).
* Number of Voicemail Messages: The total number of voicemail messages sent by the customer.
* Total Day Minutes: Total duration of calls made by the customer during the daytime (in minutes).
* Total Day Calls: Total number of daytime calls made by the customer.
* Total Day Charge: Total charges incurred by the customer for daytime calls.
* Total Evening Minutes: Total duration of calls made by the customer during the evening (in minutes).
* Total Evening Calls: Total number of evening calls made by the customer.
* Total Evening Charge: Total charges incurred by the customer for evening calls.
* Total Night Minutes: Total duration of calls made by the customer during the night (in minutes).
* Total Night Calls: Total number of nighttime calls made by the customer.
* Total Night Charge: Total charges incurred by the customer for nighttime calls.
* Total International Minutes: Total duration of international calls made by the customer (in minutes).
* Total International Calls: Total number of international calls made by the customer.
* Total International Charge: Total charges incurred by the customer for international calls.
* Customer Service Calls: Total number of calls made by the customer to customer service.
* Churn: Indicates whether the customer has terminated their contract (True/False). This is our target  feature, the primary focus of this analysis. Predicting churn is essential for understanding customer attrition and implementing effective retention strategies.

**Objectives**

* **Primary objective**

Develop a predictive model to accurately identify the key drivers of customer churn, enabling proactive measures to mitigate churn and protect revenue

* **Ancillary objectives**:

* Evaluate the impact of customer service on customer churn.
* Analyze customer behavior patterns to inform the development of strategies for reducing churn

**Metrics of Success:**

The performance of SyriaTel’s churn prediction model will be assessed using the following key metrics based on benchmarks from research studies and the telecommunications industry:

* Accuracy: (80% - 90%) indicating the proportion of correct predictions made by the model.
* Precision:(70% - 90%) reflecting the model's ability to correctly identify actual churners among those predicted to churn.
* Recall: (60% - 85%) measuring the model's effectiveness in identifying all actual churn cases.
* F1 Score:( 65%–87%) Balances precision and recall, providing a single metric for model performance.

**Load the data**

Importing essential libraries for numerical operations (NumPy), data manipulation (Pandas), data visualization (Matplotlib and Seaborn), and machine learning (Scikit-learn). Additionally, configuring notebook aesthetics ensures that visualizations are clear and easily interpretable.

**Data Preparation**

I conducted preliminary data understanding by creating a class function in separate python file, then imported it. It loads the dataset, views its contents, the dimension and a concise statistical summary.

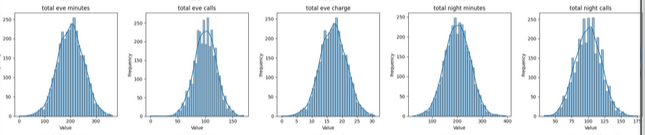
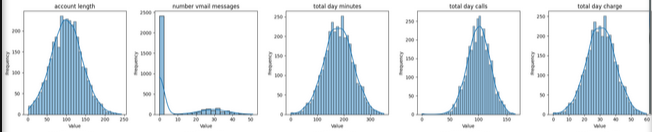
Then I proceeded to check for missing values and duplicates and found that it  didn’t have either.  
  
 **Exploratory Data Analysis (EDA)**

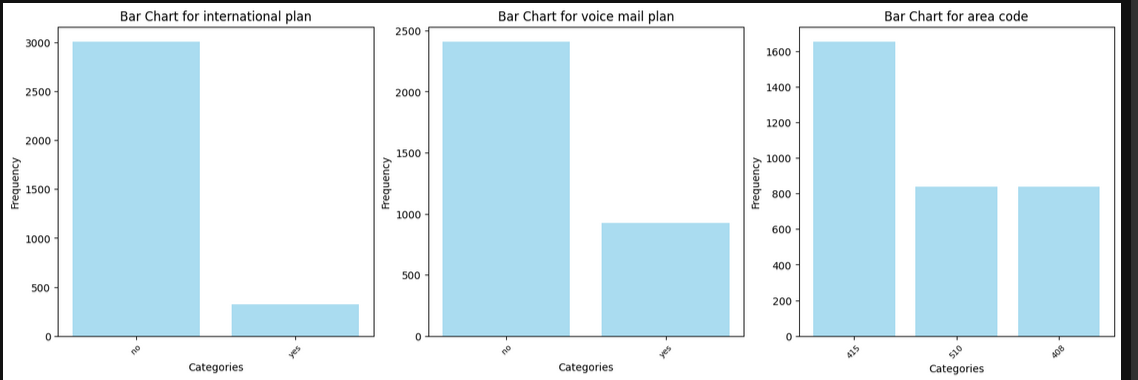
This is the process of analyzing and visualizing datasets to uncover patterns, relationships, and anomalies. It aims to understand data structure, and guide preprocessing or modeling decisions.

So here I started by identifying the categorical variables and numerical variables to determine the appropriate statistical techniques and visualizations to use for analyzing each type of data.  
 At this point I realized that state and telephone columns were irrelevant in my analysis so I drop them.  
  
 **Univariate Analysis**

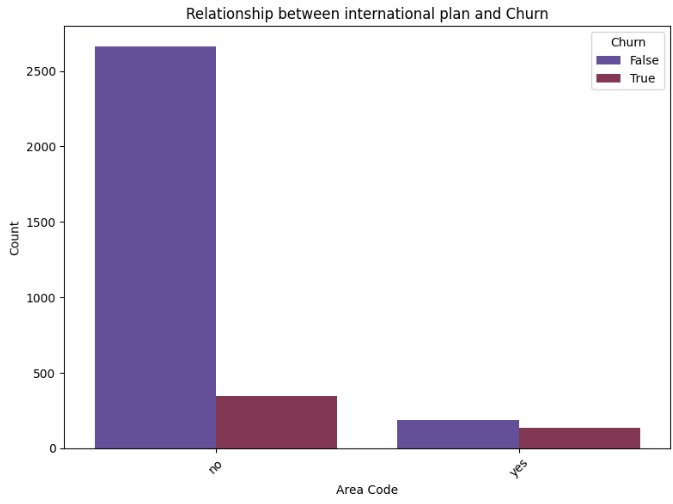
I used histograms  for univariate analysis to clearly visualize the data distribution, making it easier to identify its shape, spread, central tendency, and any outliers or patterns that may not be immediately apparent.  
 The distribution plots indicate that all features, except for customer service calls, and the number of voicemail messages, follow a normal distribution. The total number of international calls exhibits a slight right skew but remains approximately normal. The number of voicemail messages displays a pronounced peak on the right, suggesting the presence of outliers. Customer service calls exhibit multiple peaks, indicating a multimodal pattern, consistent with its nature as an integer rather than a continuous variable.

Numerical univariate

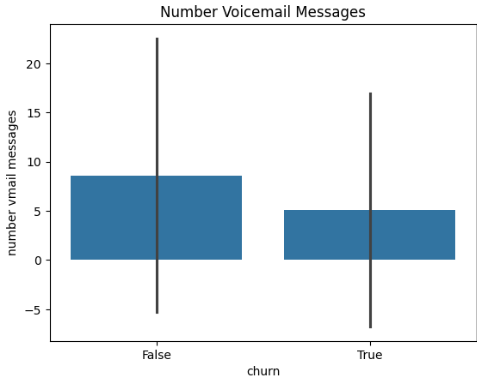




**Bivariate Analysis**

****

1. Area Code Churn Rates:
   * Area codes 408 and 510 have similar churn rates and active customer counts.
   * Area code 415 has a larger customer base, with more active and slightly more churned customers.
2. Voice Mail Plan:
   * Customers with a voice mail plan show a lower churn rate, suggesting it aids retention.
   * Most non-churning customers lack a voice mail plan, indicating it appeals to a specific segment rather than the majority.
3. International Plan:
   * International plan subscribers are fewer but have a lower churn rate.
   * Non-subscribers dominate the customer base and make up the largest group of non-churned customers.

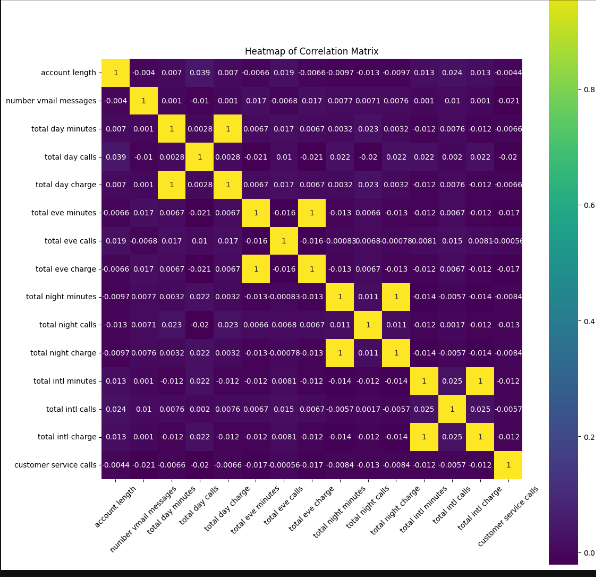
****

The density plot shows significant overlap and similar central tendencies between churn groups, indicating that account length is not a significant predictor of customer churn.

Non-churned customers tend to have higher voicemail usage compared to churned customers, who generally exhibit lower voicemail activity, possibly reflecting reduced engagement.

From the plots there seem to be a clear relationship between the number of customer service calls and the likelihood of churn. Specifically, customers who make more than four calls to customer service show a significantly higher probability of discontinuing their service.

**Multivariate Analysis**



1. Overall Correlation: Most features show little to no correlation, indicating they are mostly independent with no linear relationships.
2. Perfect Positive Correlations: Several feature pairs have a perfect positive correlation (correlation coefficient of 1), where charges are directly proportional to usage:
   * Day Minutes and Day Charges
   * Evening Minutes and Evening Charges
   * Night Minutes and Night Charges
   * International Minutes and International Charge

Handling Outliers

 Handling of outliers

Clip is a precise, non-destructive method to handle outliers by capping values within calculated bounds:

- Preserves Data: Retains all rows, avoiding data loss or bias from removal.

-  Balances Influence: Reduces outlier impact while maintaining feature relationships.

-  Interpretable: Keeps the data on its original scale, unlike complex transformations.

- Efficient and Reproducible: Systematic and computationally fast using IQR-based thresholds.

Ideal for ensuring data integrity and usability without distorting its structure or insights

**Model Building and Evaluation**

**Baseline Logistic Model**

Key findings:

* **Class 0** performed well (F1-score: 0.92, recall: 0.99).
* **Class 1** underperformed (F1-score: 0.11, recall: 0.06) due to class imbalance (2141 vs. 358 instances).

With an accuracy of 85%, the performance of class 1 needs improvement. To address this, I applied SMOTE to balance the dataset and enhance performance.

The ROC curve (AUC= 0.73**)** shows moderate model ability, but improving class 1 performance remains crucial.

**Fine-Tuning the Second Logistic Regression Model**

To improve the model, I addressed class imbalance using the SMOTE technique, scaled the data, and introduced parameters such as solver lilinea**r**, and regularization.

I selected StandardScaler because it is effective for models like Logistic Regression and Decision Trees, especially with normally distributed features. It scales data based on mean and standard deviation, offering consistency and being less sensitive to outliers compared to MinMaxScaler.

The liblinear solver is well-suited for smaller datasets and supports both **L1** and L2 regularization, making it a reliable and efficient choice for Logistic Regression optimization.

I set the **regularization parameter** to **1e12** to manage sparse features and large coefficients, ensuring robustness, stability, and reduced overfitting.

After addressing class imbalance and adding regularization, the model achieved balanced performance across both classes (0.0 and 1.0), with precision, recall, and F1-scores all at **0.76**. The overall accuracy is **76%**, with no significant difference between the classes, indicating effective handling of imbalance.

An **AUC of 0.82** demonstrates strong model performance in distinguishing between classes despite initial imbalance. The balanced metrics confirm that **SMOTE** improved the consistency and fairness of predictions across both classes.

Recommendation  
  
Focus on reducing churn by addressing frequent customer service calls (over four), as these strongly correlate with customer dissatisfaction.

Promote voicemail and international plans to specific customer segments, as both show potential to improve retention.

Develop targeted retention efforts in area code 415, leveraging its larger customer base and slightly higher churn rate.

Next steps:

Tailor marketing strategies for voicemail and international plans to high-risk groups.

Develop area-specific retention campaigns, focusing on high-churn segments.

Conclusion

he comparison of the ROC curves and their respective AUC scores clearly indicates the performance differences among the models:

1. **Baseline Model (AUC = 0.50)**:
   * This represents a random guess and sets the minimum standard for model performance.
   * An AUC of 0.50 means the model has no predictive power and is equivalent to flipping a coin.
2. **Modified Logistic Regression (AUC = 0.86)**:
   * A substantial improvement over the baseline model, this model demonstrates strong predictive ability.
   * The AUC score of 0.86 indicates that it correctly ranks positive instances above negative instances 86% of the time.
3. **Random Forest Model (AUC = 0.92)**:
   * The best-performing model in this comparison, with an AUC score of 0.92.
   * This high score signifies excellent predictive power, likely due to the Random Forest's ability to handle complex interactions and nonlinear relationships effectively.
4. **Decision Tree Model (AUC = 0.89)**:
   * Performs slightly worse than the Random Forest but still better than Logistic Regression.
   * The AUC of 0.89 suggests robust predictive capabilities, though it may lack some of the ensemble advantages that the Random Forest leverages.

**Compelling Conclusion:**

The **Random Forest model (AUC = 0.92)** emerges as the most effective model, delivering the highest predictive accuracy among all candidates. It is particularly suitable for this problem due to its capacity to manage complex patterns in the data.

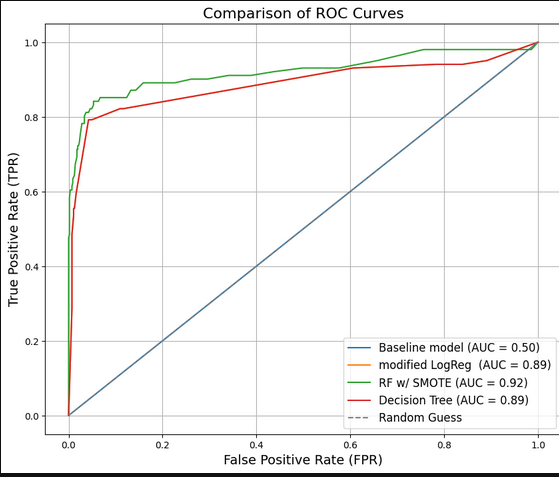
The **Decision Tree model (AUC = 0.89)** also performs well, but its slightly lower AUC indicates it may not generalize as effectively as the Random Forest.

The **Modified Logistic Regression (AUC = 0.86)** is a strong, simpler alternative. It offers solid performance with less computational overhead, making it a viable option when interpretability or efficiency is prioritized over raw predictive power.

Finally, the **Baseline Model (AUC = 0.50)** serves as a reminder of the stark contrast between meaningful predictions and random guessing. The significant improvement from the baseline underscores the added value of employing these sophisticated models.

In summary:

* If **maximizing predictive performance** is critical, go with the **Random Forest**.
* If you value **interpretability and simplicity**, the **Modified Logistic Regression** is a practical choice.
* The **Decision Tree** strikes a middle ground, offering competitive accuracy with more straightforward interpretability than Random Forest.



1.       **Research References:**

[Productplan.com](https://www.productplan.com/glossary/churn/#:~:text=Definition%3A%20Churn%20is%20a%20measurement,with%20a%20product%20or%20service.)

·       [Customer churn prediction in telecom using machine learning](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6)

·       [Reducing Customer Churn: Using Machine Lear](https://github.com/liusseka/DSF-Phase-3-Project-Predicting-Customer-Churn-at-SyriaTel-with-Machine-Learning)

[Worldfolio](https://www.theworldfolio.com/company/syriatel-syria/197/)

[Kaggle](https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset/discussion/448926)

References: https://medium.com/@cmfritz0/predicting-customer-churn-a2df316ced81