**Customer churn prediction: Telecom Churn Dataset**

# **1. Overview:**

* Customer churn, also referred to as customer attrition or turnover, is the phenomenon where customers cease using a company's products or services. It is a critical metric for businesses, especially in industries like telecommunications, internet services, insurance, and subscription-based models, where retaining existing customers is significantly more cost-effective than acquiring new ones. Companies often analyze churn to identify at-risk customers and implement targeted retention strategies.
* Churn can be categorized into voluntary churn, where customers actively switch to competitors due to dissatisfaction or better offers, and involuntary churn, which occurs due to external factors like relocation or life events. Predictive analytics plays a crucial role in addressing voluntary churn by using machine learning models to identify customers most likely to churn. These models enable businesses to prioritize retention efforts, improve customer satisfaction, and ultimately reduce revenue loss. This project focuses on predicting customer churn in the telecom industry using a dataset that captures customer demographics, service usage, and contract details, aiming to provide actionable insights for customer retention strategies.

# **2. Introduction:**

* **Objective:** The goal of this project is to predict customer churn for a telecom company using machine learning techniques. Customer churn refers to the phenomenon where customers stop using a company's services. Predicting churn helps businesses take proactive measures to retain customers.
* **Dataset:** The dataset used is the Telecom Churn Dataset, which contains customer information, usage patterns, and whether the customer churned or not.
* **Approach:** The notebook follows a structured machine learning workflow, including data exploration, preprocessing, feature engineering, model training, evaluation, and interpretation.

# **3. Data Exploration:**

* **Distribution of Churn and Non-Churn Customers:**Churn analysis is at the heart of our exploration, as it directly pertains to the primary objective of understanding and reducing customer attrition.
* **Churn Status:  
  1-False (Didn't Churn):** 2,850 customers **2-True (Churned):** 483 customersFrom the preliminary data, we can observe that a vast majority of customers have chosen to stay with the service, whereas a smaller fraction have decided to churn. This distribution sets

the stage for a deeper dive into understanding the reasons behind these decisions.

* **Service Plans Distribution:  
  1-Voice Mail Plan:  
  Subscribed (Yes):** 922 customers **Not Subscribed (No):** 2,411 customers

**2-International Plan:**  
**Subscribed (Yes):** 323 customers  
**Not Subscribed (No):** 3,010 customers  
The data suggests a larger preference towards not subscribing to both the Voice Mail and International plans. The reasons for these choices could be manifold - cost considerations, perceived value, or other alternatives. A more in-depth analysis could reveal insights about customer preferences.

* **State-wise Distribution:**The dataset spans customers from 51 unique states, such as 'KS', 'OH', and 'NJ', to name a few. Given the broad geographical coverage, understanding state-wise churn trends could provide insights into regional preferences or challenges.
* **Total International Records  
  1-Total intl minutes:** This metric provides insights into the duration of international calls. Specifically, it represents the cumulative minutes a customer has spent on overseas calls. **2-Total intl calls:** This signifies the frequency of international interactions, reflecting the total number of international calls initiated by a customer. **3-Total intl charge:** A representation of the financial aspect, this denotes the total expenses or charges borne by a customer due to their international call usage

# **4. Feature Selection and Engineering:**

* **Identifying Relevant Features**Recognizing the relevance of each feature ensures a streamlined model that's both efficient and effective.  
  **1-Statistical Tests:** Use statistical tests to determine the relationship between each feature and the target variable. For instance, the chi-squared test can be used for categorical features.
* **State vs. Churn:** P-value of 0.0023 suggests that the customer's state has a statistically significant association with churn status.
* **International plan vs. Churn:** An extremely low p-value (2.49e-50) shows a strong association between having an international plan and churning.
* **Voice mail plan vs. Churn:** The p-value (5.15e-09) also suggests a strong relationship between having a voice mail plan and the likelihood to churn. In summary, both the international plan and voice mail plan have pronounced associations with churn, while regional differences (State) also seem to play a role.

**2-Domain-specific Features:** Incorporating domain knowledge aids in refining feature selection. For example, if customers making a high number of service calls correlate with higher churn, one could introduce a binary feature to capture this relationship.  
**3-Binning:** In future works, a notable strategy is binning, such as converting continuous variables like the State into categorical ones. This can unveil non-linear relationships that might otherwise remain obscured.

* **Anomaly Detection:**  
  Uncovering anomalies is essential, as outliers can influence model performance and mislead predictive patterns.  
  **1-IQR (Interquartile Range):** A popular statistical approach to identify outliers. It determines data points by gauging the dispersion of the data. Anomalies are those that fall outside the established range, typically demarcated by 1.5 times the IQR above the third quartile or below the first quartile.  
  **2-Findings:** Based on the IQR method, we identified 529 rows in our dataset containing outliers.
* **Correlation and Feature Importance**  
  Deciphering inter-feature relationships and their correlation with 'Churn' is fundamental for informed modeling.

**1-Correlation Matrix:** This visual tool elucidates linear associations between features. Notably, features manifesting robust correlations with 'Churn' become central to the modeling process. Conversely, those exhibiting high mutual correlations might be redundant, offering avenues for model simplification.  
**2-Feature Importance Plot:** Advanced algorithms, like Random Forest, bestow the capability to rank features by importance. Visualizing these rankings crystallizes which attributes significantly influence churn predictions, informing strategy and intervention pathways.

**3-Key Findings:**

* **International Calls:** Demonstrates the strongest positive correlation with 'Churn' at 0.22. This suggests that as the number of international calls increases, the likelihood of churn also rises.
* **Customer Service Calls, Total Day Minutes & Total Day Charge:** Each of these features exhibits a noteworthy positive correlation of around 0.2 with 'Churn'. This indicates that factors like longer daily usage or increased customer service interactions might elevate churn propensity.
  + **Voice Mail Plan:** This feature shows a negative correlation of -0.1 with 'Churn'. This implies that subscribers with a voice mail plan are slightly less likely to churn compared to those without one.These correlations offer a guiding framework for the predictive modeling phase, highlighting which features might be especially influential in determining customer churn.

# **5.Data Preprocessing:**

* **Data preprocessing**: plays a pivotal role in preparing the data for effective modeling. In this section, we address class imbalance with oversampling and delineate our data into training and testing sets.
* **Oversampling:  
  Objective:** Address class imbalance, which might otherwise influence model performance unfavorably.  
  **Technique Used:**  
  **SMOTE (Synthetic Minority Over-sampling Technique):** This technique creates synthetic instances of the minority class by considering the feature space, providing a balanced representation.

**Outcome:** Post-SMOTE application, the 'Churn' feature distribution is:  
No Churn: 2137 samples  
Churn: 2137 samples

* **Train-Test Split:**

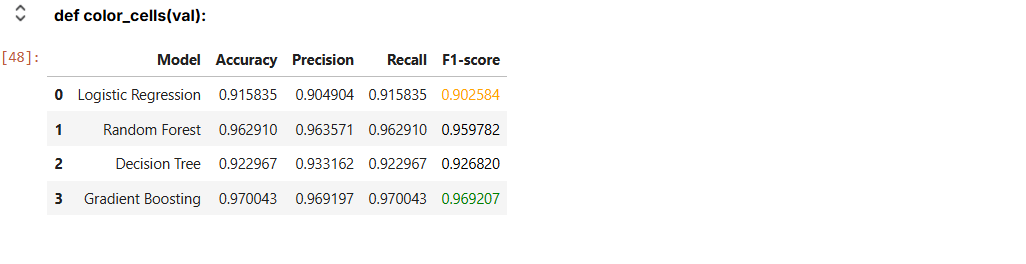
**Objective:** Ensure model generalization by training on one data subset and validating on another.  
Procedure:  
1-Opt for a 75-25 split. 75% of the data is earmarked for training, with the remaining 25% for validation.  
2-Implement train\_test\_split for randomized partitioning.  
3-Invoke the stratify parameter to maintain consistent 'Churn' distribution in both training and test datasets.

* **Colinearity problem:**

Several features have high correlation between them which are the following features **(Total day minutes, Total eve minutes, Total night minutes, Total intl minutes , Total day charge, Total eve charge, Total night charge, Total intl charge)** will be reduced to four features, the first four features will be dropped

# **6. Model Development:**

* **Pipeline:**  
  The pipeline serves to automate the end-to-end process, orchestrating the flow from preprocessing stages to model training, ensuring uniform handling of data at every stage.
* **RobustScaler**:  
  It scales features to diminish the impact of outliers. By working based on the median and interquartile range, it offers robustness against outlier effects.  
  **1-Logistic Regression:** This is a linear technique tailored for binary classification tasks, modeling the likelihood of a sample fitting into a designated class.  
  **2-Random Classifier:** Acting as a baseline classifier, its predictions are random, thereby setting a foundational performance metric.  
  **3-Decision Tree Model:** A transparent model that makes decisions reliant on feature values, crafting a tree-structured decision schema.  
  **4-Gradient Boosting:** An advanced ensemble methodology that progressively crafts trees. Each new tree seeks to amend the shortcomings of the prior one, often yielding enhanced accuracy.  
  Note: A meticulous grid search was executed for the Random Forest and Gradient Boosting models to pinpoint the best hyperparameters. These optimal parameters are now integral to the pipeline, priming the models for peak performance.
* **Final Model Selection:** After evaluating the performance metrics of all models, the best-performing model will be chosen. This selected model will form the basis for our subsequent analyses and predictions moving forward. **(The best model is Gradient Boosting with an accuracy of 0.97)**



# **7. Model Evaluation:**

* **Evaluating model performance** is crucial to ensure our predictions are both accurate and reliable. We employ a series of metrics and techniques to critically assess how well our models have been trained and how they might perform on unseen data.
* **Metrics and Tools:**  
  **1-Classification Report:** This provides a comprehensive breakdown of performance metrics for each class, including precision, recall, f1-score, and support. It's a quick way to gauge overall model performance across categories.  
  **2-Confusion Matrix:** A table used to describe the performance of a classification model on a set of data for which the true values are known. It contrasts actual versus predicted classifications.  
  **3-Precision-Recall Curve:** A graphical representation that plots precision against recall for different thresholds. This curve showcases the trade-off between precision and recall for different threshold values and is particularly useful when classes are imbalanced.
* **Performance Summary:**

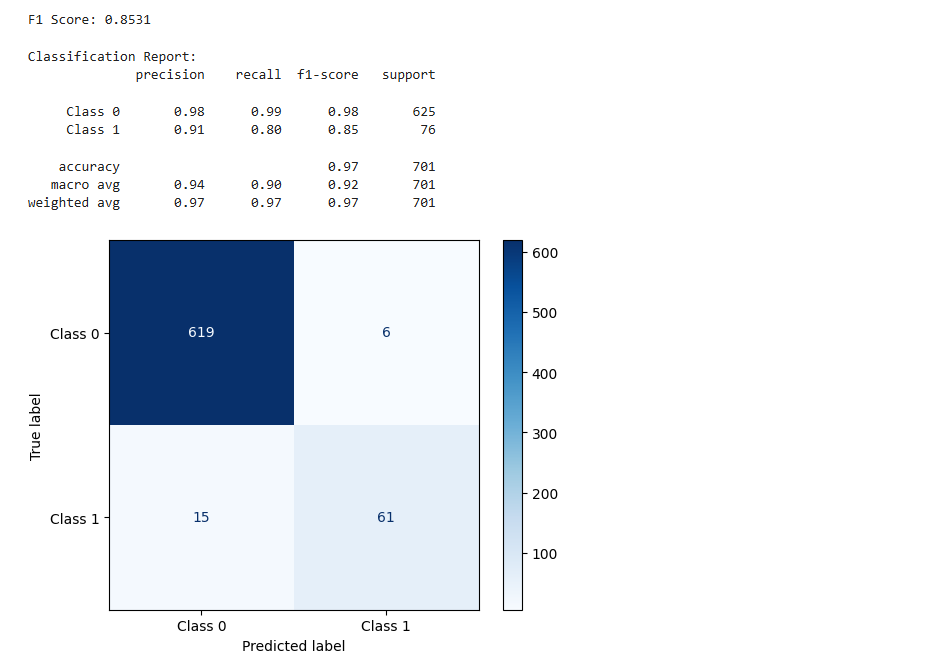
Post evaluation, we collate the critical metrics into a new dataframe for a side-by-side comparison. This dataframe lists:

**1-Accuracy:** The ratio of correctly predicted instances to the total instances.

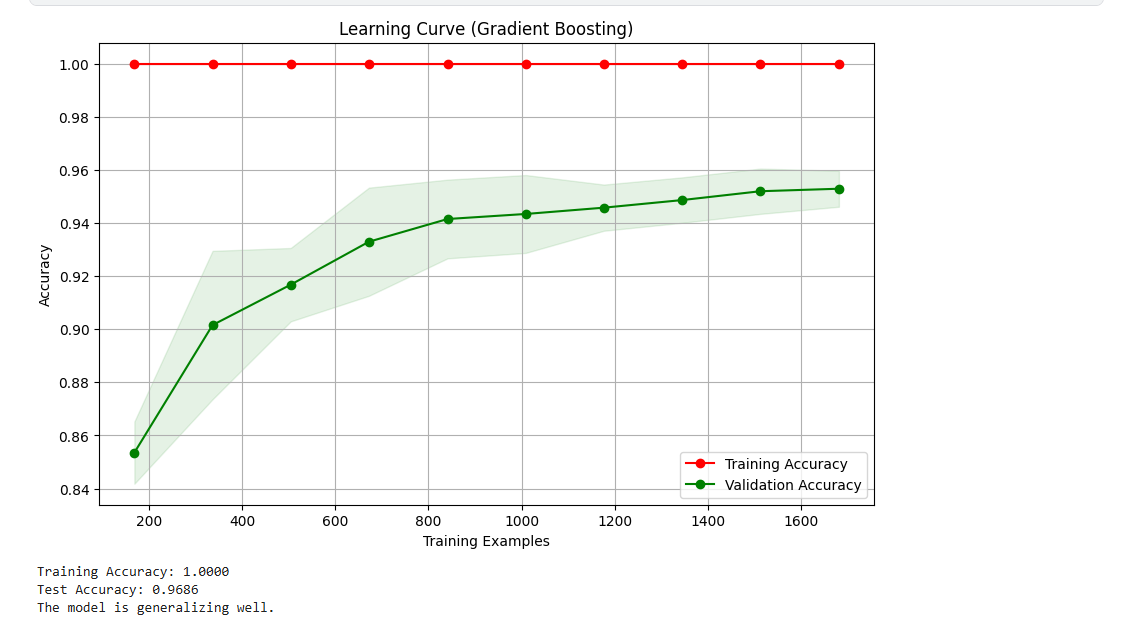
**2-Precision:** The ratio of correctly predicted positive observations to the total predicted positives.

**3-Recall (Sensitivity):** The ratio of correctly predicted positive observations to all the observations in the actual class.

**4-F1 Score:** The weighted average of precision and recall.



* **Using function to plot learning curve:  
  to check about overfitting**

****

# **8. Conclusion:**

* **Summary of Findings:** The project successfully built a predictive model for customer churn with reasonable accuracy and recall. The results can help the telecom company identify at-risk customers and implement retention strategies.
* **Limitations:**  
  The dataset had class imbalance, which affected model performance.  
  External factors (e.g., market trends, competitor actions) were not included in the dataset.
* **Future Work:**  
  Incorporate additional data sources (e.g., customer feedback, social media activity).  
  Experiment with deep learning models for improved performance.

# **9. Appendix:**

**Notebook Link:** [**https://www.kaggle.com/code/ahmedelhamamy/dsmentorship-task**](https://www.kaggle.com/code/ahmedelhamamy/dsmentorship-task)