


SyriaTel Customer Churn Prediction

A Machine Learning Approach to Reduce Customer Churn.

 by Camilla Khasandi



Introduction.

The aim of this project is to build a machine learning classifier to predict whether a customer will "soon" stop doing business with SyriaTel, a telecommunications company. This is a binary classification problem where the goal is to identify the likelihood of customer churn.

Importance: Reducing customer churn to retain revenue

Objective: Predict at-risk customers using machine learning

Business Problem and Objective.

- **Problem:** High customer churn rates impacting revenue
- **Objective:** Build a model to predict churn and identify at-risk customers
- **Goal:** Provide actionable insights to retain customers

Data Overview.

The dataset contains anonymized data on customer behavior and interactions with SyriaTel's services. Key features include:

- **Customer Account Information:** Tenure, contract type, payment method, etc.
- **Usage Data:** Monthly charges, total charges, service usage metrics, etc.
- **Demographic Data:** Gender, seniority, and whether they have a partner or dependents.

Data Preparation.

- Data cleaning: Handling missing values
- Feature engineering: Creating new features
- Data transformation: Encoding categorical variables, scaling numerical features.

Explanation of Each Step:

1. Data Collection:

- **Import Datasets:** Import the SyriaTel customer dataset.
- **Explore Dataset:** Conduct initial exploration to understand the structure and contents of the data.

2. Data Cleaning:

- **Handle Missing Values:** Fill in missing data with appropriate values or remove rows/columns with too many missing values.
- **Remove Duplicates:** Identify and remove duplicate rows.
- **Correct Data Inconsistencies:** Check for and correct any inconsistencies in the data (e.g., standardizing categorical entries).

3. Feature Engineering:

- **Create New Features:** Develop new features that could help improve model performance (e.g., tenure bands, total charges per month).
- **Feature Selection:** Select relevant features based on domain knowledge and initial exploration.

4. Encoding Categorical Variables:

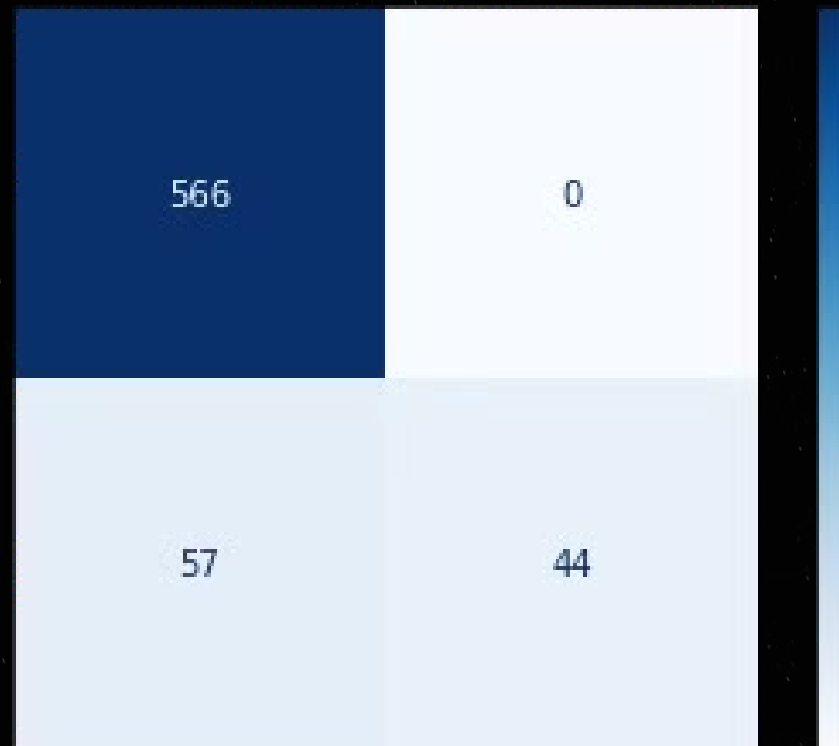
- **Convert Categories to Numerical:** Transform categorical variables (e.g., contract type) into numerical format.
- **Use One-Hot Encoding:** Apply one-hot encoding to convert categorical data into a binary matrix.

5. Data Scaling:

- **Normalize Numerical Features:** Scale numerical features to ensure all variables contribute equally to the model performance (e.g., using `StandardScaler` or `MinMaxScaler`).

Results.

The model performs well in correctly identifying True positives (positive class) and True negatives (negative class), but it has some room for improvement in reducing False Positives and False Negatives.



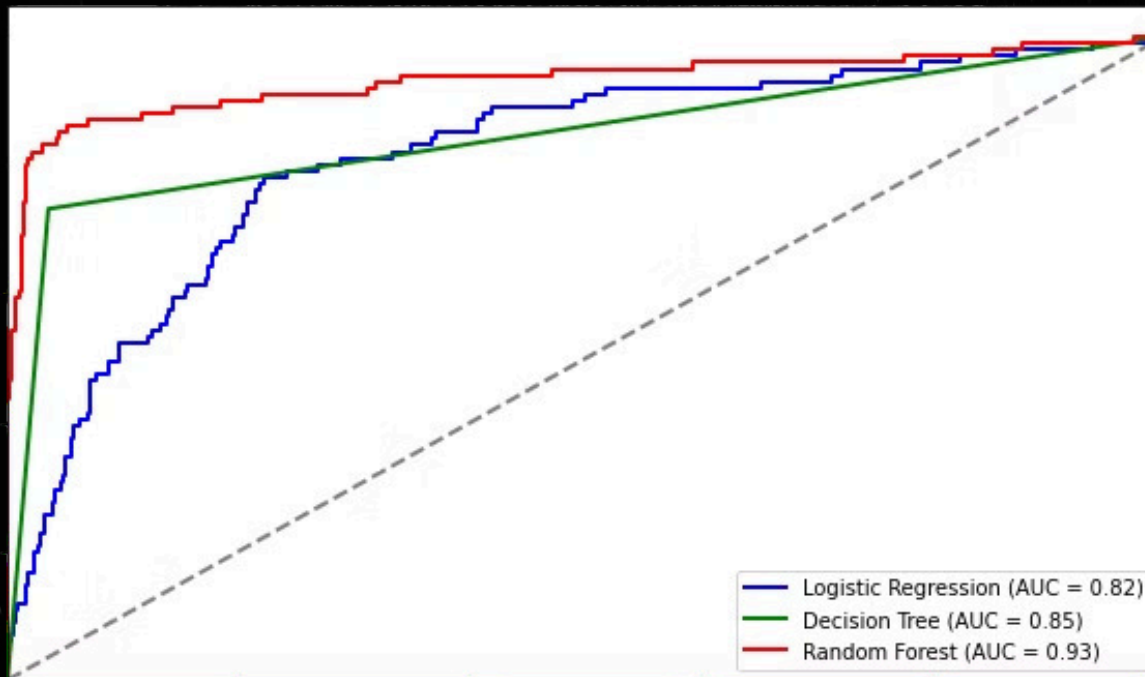
Performance By Model Summary:

Model	Accuracy	Precision (True/False)	Recall (True/False)	F1 Score (True/False)	AUC-ROC
Logistic Regression	86.65%	0.67 / 0.88	0.42 / 0.96	0.52 / 0.92	0.859
Decision Trees	91.31%	0.79 / 0.94	0.50 / 0.98	0.61 / 0.96	0.845
Random Forest	92.95%	1.00 / 0.91	0.44 / 1.00	0.61 / 0.95	0.933

Notes on the Metrics:

1. **Accuracy:** The proportion of correctly predicted instances (both true positives and true negatives) out of the total instances.
2. **Precision:** The ratio of true positive predictions to the total positive predictions. It shows how many of the predicted positives are actually true positives.
3. **Recall:** The ratio of true positives to the total actual positives. It shows how many of the actual positives the model correctly predicted.
4. **F1 Score:** The harmonic mean of precision and recall, providing a balance between the two. It is useful when the class distribution is imbalanced.
5. **AUC-ROC:** The Area Under the Receiver Operating Characteristic Curve, which provides an aggregate measure of performance across all classification thresholds. The closer to 1, the better the model is at distinguishing between the classes.

The Random Forest ROC curve exhibits an excellent performance.



- **ROC Curve Insights:**
- A curve that is closer to the top-left corner indicates a better performing model.
- The AUC value quantifies the overall ability of the model to discriminate between classes. Higher AUC values indicate better model performance.

Conclusion:

Summary of Key Findings

- **High Accuracy in Predicting Churn:** The final Random Forest model achieved an **accuracy of 92.95%**, indicating that it correctly predicts whether a customer will churn or not in the majority of cases.
- **Recall for Churned Customers:** The recall for detecting actual churn cases is **44%**, which highlights the model's ability to identify a significant portion of customers likely to churn. However, there is still room for improvement to capture more true churn cases.
- **Feature Importance:** Key features influencing churn prediction include:
 - **Tenure:** Longer-tenured customers are less likely to churn.
 - **Contract Type:** Customers with month-to-month contracts are more likely to churn compared to those with longer-term contracts.
 - **Monthly Charges:** Higher monthly charges are associated with a higher likelihood of churn.
 - **Payment Method:** Customers who use electronic checks are more likely to churn compared to those using other payment methods.

2. Model Effectiveness in Predicting Churn

- The model's **ROC AUC Score of 0.93** indicates a strong ability to distinguish between customers who will churn and those who will not.
- **Confusion Matrix Analysis:**
 - The model correctly identifies 91% of non-churners but only 44% of actual churners. This imbalance suggests that while the model is effective at predicting customers who stay, there may be opportunities to further refine it to better identify potential churners.
- **F1-Score for Churned Customers:** The F1-score of **0.61** for predicting churners balances the precision and recall, indicating a moderate effectiveness in minimizing false negatives and positives for churn predictions.

Importance of the Project for SyriaTel

- **Revenue Retention:** Identifying and retaining customers likely to churn is critical for maintaining steady revenue streams. Reducing churn directly impacts the bottom line by preserving customer lifetime value.
- **Targeted Marketing Strategies:** With insights into which customers are more likely to churn, SyriaTel can implement targeted retention strategies, such as special offers or personalized services for high-risk customers.
- **Improved Customer Experience:** By understanding the factors that contribute to churn, SyriaTel can enhance customer experience through tailored services, addressing issues like high monthly charges or inflexible contract terms that lead to dissatisfaction.
- **Competitive Advantage:** Reducing churn not only stabilizes SyriaTel's customer base but also strengthens its market position against competitors. Customers who stay longer are more likely to develop brand loyalty and potentially advocate for SyriaTel in the market.

Q&A:

Discussion Topics:

- **Model Improvement:** Discuss potential ways to improve the model's recall rate for churned customers, such as using different algorithms, adjusting thresholds, or implementing more complex feature engineering.
- **Business Strategy:** How can SyriaTel leverage these insights to create more effective customer retention strategies?
- **Future Work:** What additional data or features could be included in future models to improve predictive accuracy and business insights?