Syriatel Churn Customer Prediction

Project Overview

Customer churn occurs when customers stop using a business or service. It’s especially important in industries like telecom, where retaining customers is cheaper than acquiring new ones. Churn can be detected by analyzing behaviors like reduced usage, complaints, or canceled subscriptions. Machine learning models use factors such as customer care calls, service plans, and billing data to identify at-risk customers.

Problem statement

Telecom customer churn can cause major revenue losses. Understanding why customers leave and predicting those at risk helps providers take action to retain them. This project aims to predict customer churn using a machine learning model.

Proposed Solution (Analysis & Modelling) and Projected Conclusion

This project proposes using a machine learning model to predict customer churn based on client characteristics and behaviors. Decision trees or logistic regression will be used for their effectiveness and interpretability in churn prediction. Key factors like service usage, plan details, and customer complaints will be analyzed. The process includes data preprocessing, model optimization, and performance evaluation using metrics such as ROC-AUC and F1-score.

The goal is to develop a robust churn prediction system to help stakeholders identify and engage at-risk clients proactively. By leveraging the model's insights, the business can reduce churn, enhance customer satisfaction, and retain revenue through targeted interventions.

Objectives

* Identify the factors that lead to customer churn.
* Accurately predict which customers are at risk of churning.
* Take proactive steps to retain customers who are at risk of churning.

Metrics of success

- Accuracy Score: ≥ 80%
- Recall Score: ≥ 69%

DATA UNDERSTANDING

- **Source of Data:** Kaggle
- **Data Description:** The dataset contains 3333 rows and 21 columns, including demographic data, usage statistics, service plans, and the churn target variable.

Numeric Columns:

1. **account length:** The number of days or months a customer has been subscribed to the service.
2. **area code:** A numeric code representing the geographical area where the customer's phone is registered.
3. **number voice mail messages:** The count of voice mail messages stored in the customer's account.
4. **total day minutes:** The total number of minutes the customer used during the day.
5. **total day calls:** The total number of calls made during the day.
6. **total day charge:** The total cost of calls made during the day.
7. **total eve minutes:** The total number of minutes the customer used during the evening.
8. **total eve calls:** The total number of calls made during the evening.
9. **total eve charge:** The total cost of calls made during the evening.
10. **total night minutes:** The total number of minutes the customer used during the night.
11. **total night calls:** The total number of calls made during the night.
12. **total night charge:** The total cost of calls made during the night.
13. **total intl minutes:** The total number of international minutes used by the customer.
14. **total intl calls:** The total number of international calls made by the customer.
15. **total intl charge:** The total cost of international calls.
16. **customer service calls:** The total number of times the customer called customer service.

Categorical Columns:

1. **state:** The state where the customer resides.
2. **phone number:** The customer's unique phone number.
3. **international plan:** Indicates whether the customer has subscribed to an international call plan.
4. **voice mail plan:** Indicates whether the customer has subscribed to a voicemail plan .

DATA PREPARATION & ANALYSIS

Data Preparation

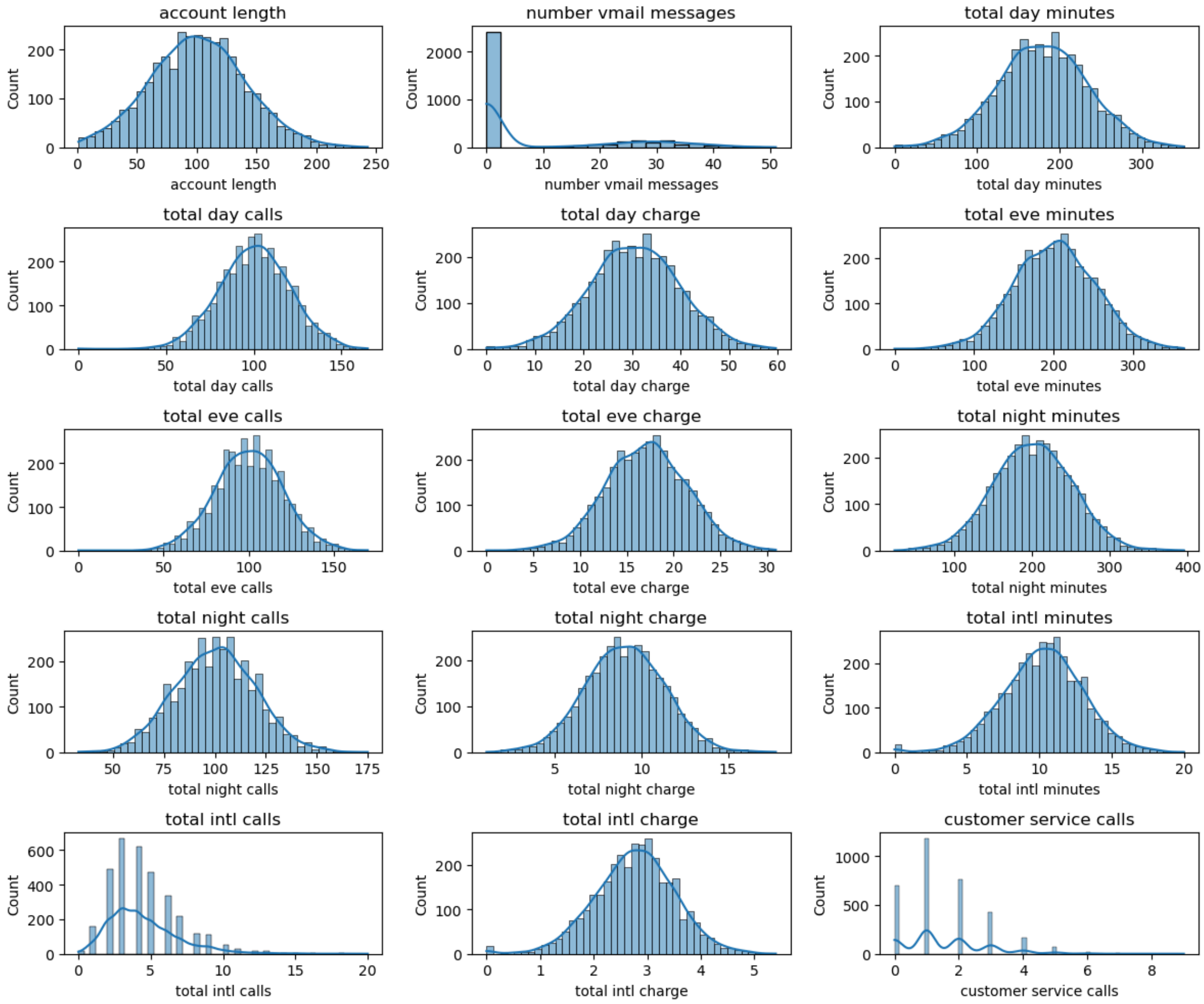
- **Checks Performed:**
  - No missing or null values.
  - No duplicate rows detected.
  - Outlier analysis on numeric columns

Actions Taken:

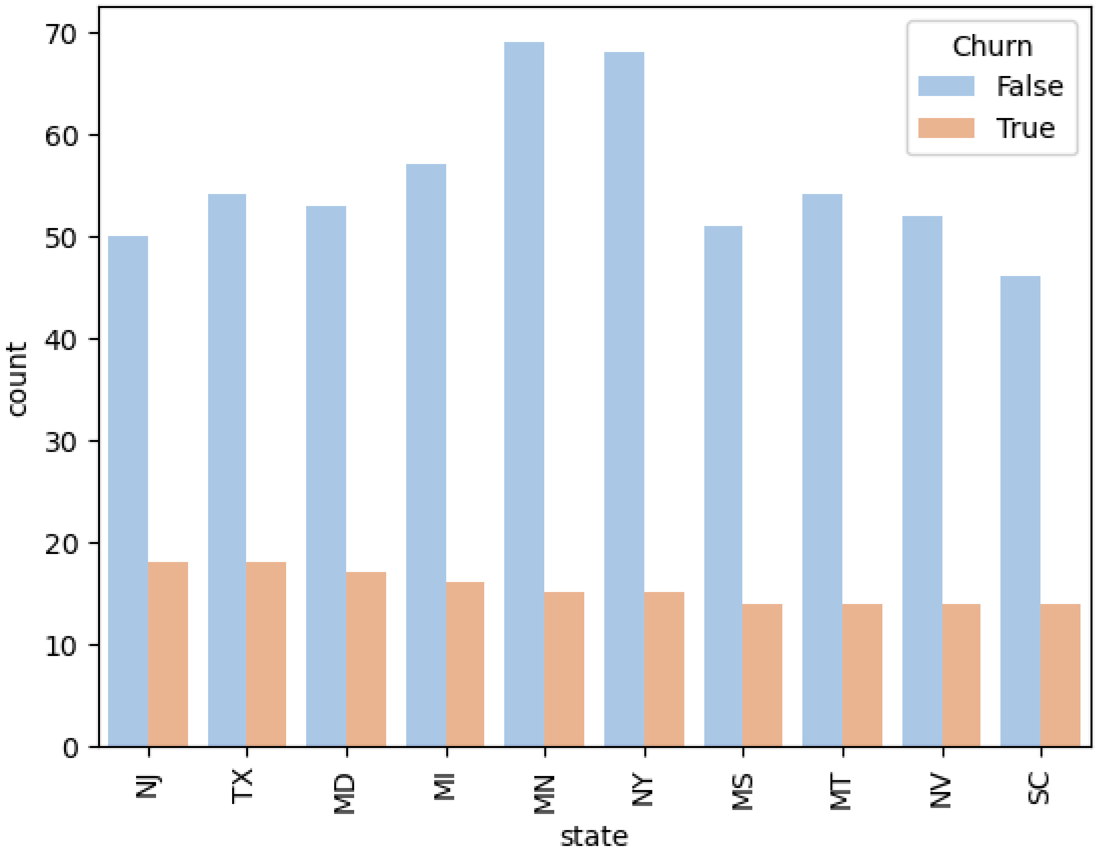
- Encoded categorical features (state, international_plan, etc.) using one-hot encoding.

Data Analysis

- **Univariate Analysis**:
  - Distribution of churn: Approximately 80% non-churners and 15% churners (class imbalance).
  - Usage patterns of services (for example; total day minutes distribution).



  -
- **Bivariate & Multivariate Analysis**:
  - Churn rates are higher for customers with the international_plan.
  - High correlation between total day charges and total day minutes.



Modelling

Models Used

Logistic Regression.

- Justification: Offers interpretability and works well for linear relationships.

Metrics of Success

- Focused on **ROC-AUC** >= 80% and **Recall**>= 69% to ensure fewer false negatives.

Evaluation

Three models—Baseline Decision Tree, Hyperparameter-Tuned Decision Tree, and Logistic Regression—were evaluated for predicting customer attrition.

- **Baseline Decision Tree:** Achieved 100% accuracy but showed signs of overfitting, limiting its generalizability.
- **Logistic Regression:** Delivered consistent results with 65% accuracy and 73% recall, despite struggling with non-linear patterns.
- **Hyperparameter-Tuned Decision Tree:** Emerged as the best model with 87% accuracy, 75% recall, and a ROC-AUC of 80%, addressing overfitting issues while balancing precision and recall effectively.

The hyperparameter-tuned Decision Tree is the most suitable for deployment due to its ability to capture non-linear interactions and proactively identify at-risk customers.

Conclusion:

Model Performance:

The Logistic Regression classifier performed decently with an AUC of 0.6899, indicating that the model has moderate discriminative power. The ROC Curve analysis shows that the model is able to distinguish between the two classes, but there is still potential for improvement.

Feature Importance:

The feature importance analysis shows which features most significantly contribute to the predictions made by the Logistic Regression model. Understanding these important features helps in interpreting the model’s decision-making process and can guide further improvements in feature selection or engineering.

Data Preprocessing:

Scaling the data and applying SMOTE (Synthetic Minority Oversampling Technique) helped address the class imbalance and scale the features, which improved the Logistic Regression model's performance.

Recommendations:

Model Enhancement: Hyperparameter Tuning: Perform hyperparameter tuning using techniques like Grid Search or Random Search to improve the Logistic Regression model. This could help find the optimal regularization strength or solver for the model. Feature Engineering: Interaction Features: Consider creating interaction terms between important features to capture more complex relationships in the dataset. Feature Selection: After analyzing the feature importance, remove less relevant features to avoid overfitting and make the model more efficient. Evaluation Metrics: In addition to AUC and the ROC Curve, assess other metrics like precision, recall, and F1-score, especially considering the class imbalance in the dataset. These metrics will give a more complete picture of model performance. Model Interpretability: Examine Coefficients: Carefully examine the coefficients of the Logistic Regression model to understand the impact of each feature on the prediction. Features with very large coefficients might need further scrutiny or scaling adjustments. By continuing to refine the Logistic Regression model through these steps, you can further improve its accuracy and interpretability.

Click to add a cell.