SYRIATEL CUSTOMER CHURN PROJECT REPORT

Project Overview

When customers discontinue using a business or service, it's known as customer churn. It is a crucial indicator in sectors like telecoms, where keeping current clients is frequently more economical than finding new ones. By examining patterns of client behavior, such as decreased service usage, frequent complaints, or canceled subscriptions, churn can be identified. The amount of customer care calls, service plans, and billing information are just a few examples of the characteristics that machine learning models might use to detect at-risk customers.

Problem statement

Telecom customer churn is a critical issue that can lead to significant revenue loss for service provider institutions. Understanding why customers leave the telecom service provider and predicting which customers are at risk of churning enables the telecom service provider to take proactive measures to retain customers. In this project I aim to predict customer churn using a machine learning model.

Proposed Solution (Analysis & Modelling) and Projected Conclusion

I suggest creating a machine learning model with a dataset of client characteristics and behaviors in order to forecast churn. Because they are effective and interpretable for churn prediction tasks, decision trees or logistic regression models will be used. The effect of important elements on turnover, such as service usage, plan specifics, and customer complaints, will be examined. Preprocessing the data, optimizing the models, and assessing performance using metrics like ROC-AUC and F1-score are all part of the project.

A strong churn prediction system that can assist stakeholders in proactively identifying and interacting with at-risk clients is the anticipated result of this project. The business can lower churn rates, increase overall customer satisfaction, and retain revenue by utilizing the model's findings to perform focused interventions.

Objectives

- Develop a machine learning model to predict customer churn.
- Achieve high accuracy and recall to minimize false negatives (missed churners).
- Provide actionable insights for stakeholders to reduce churn rates.

Metrics of success

- Accuracy Score: $\geq 85\%$
- Recall Score: $\geq 80\%$

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 vmail 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 (e.g., "Yes" or "No").
4. **voice mail plan:** Indicates whether the customer has subscribed to a voicemail plan (e.g., "Yes" or "No").

DATA PREPARATION & ANALYSIS

Data Preparation

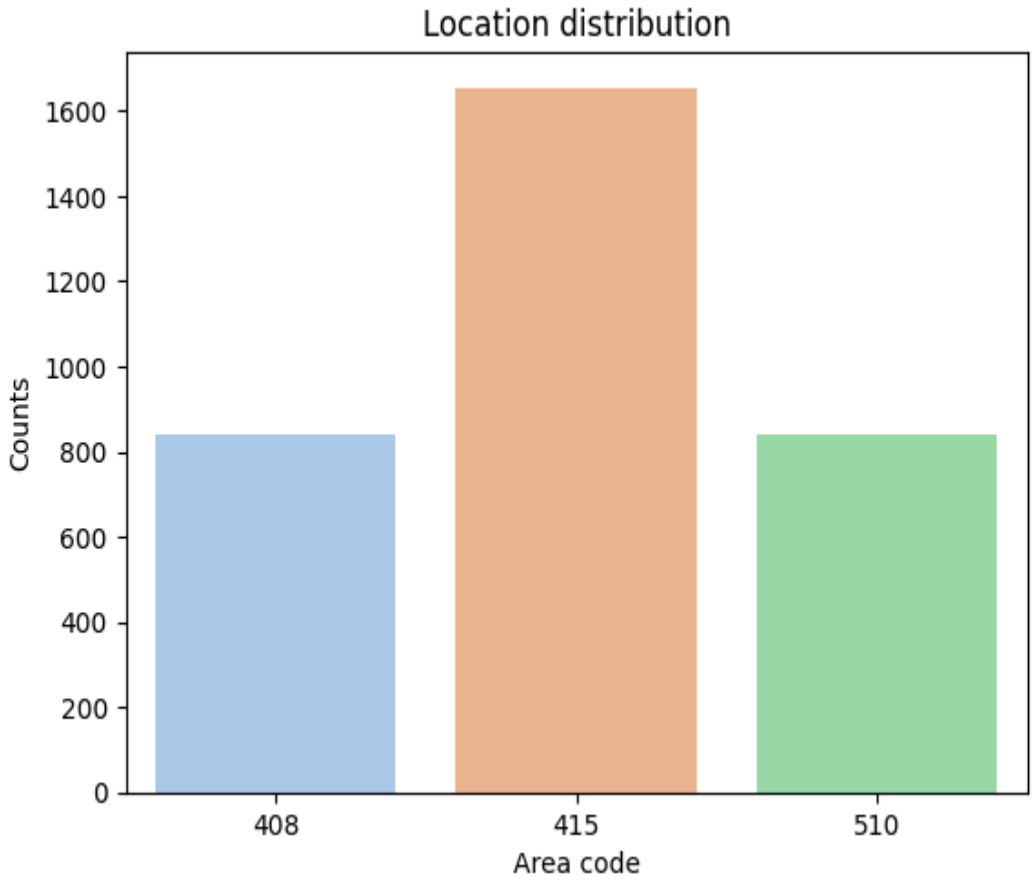
- **Checks Performed:**
    - No missing or null values.
    - No duplicate rows detected.
    - Outlier analysis on numeric columns

Actions Taken:

- Dropped irrelevant columns such as phone number column.
- Encoded categorical features (state, international_plan, etc.) using one-hot encoding.

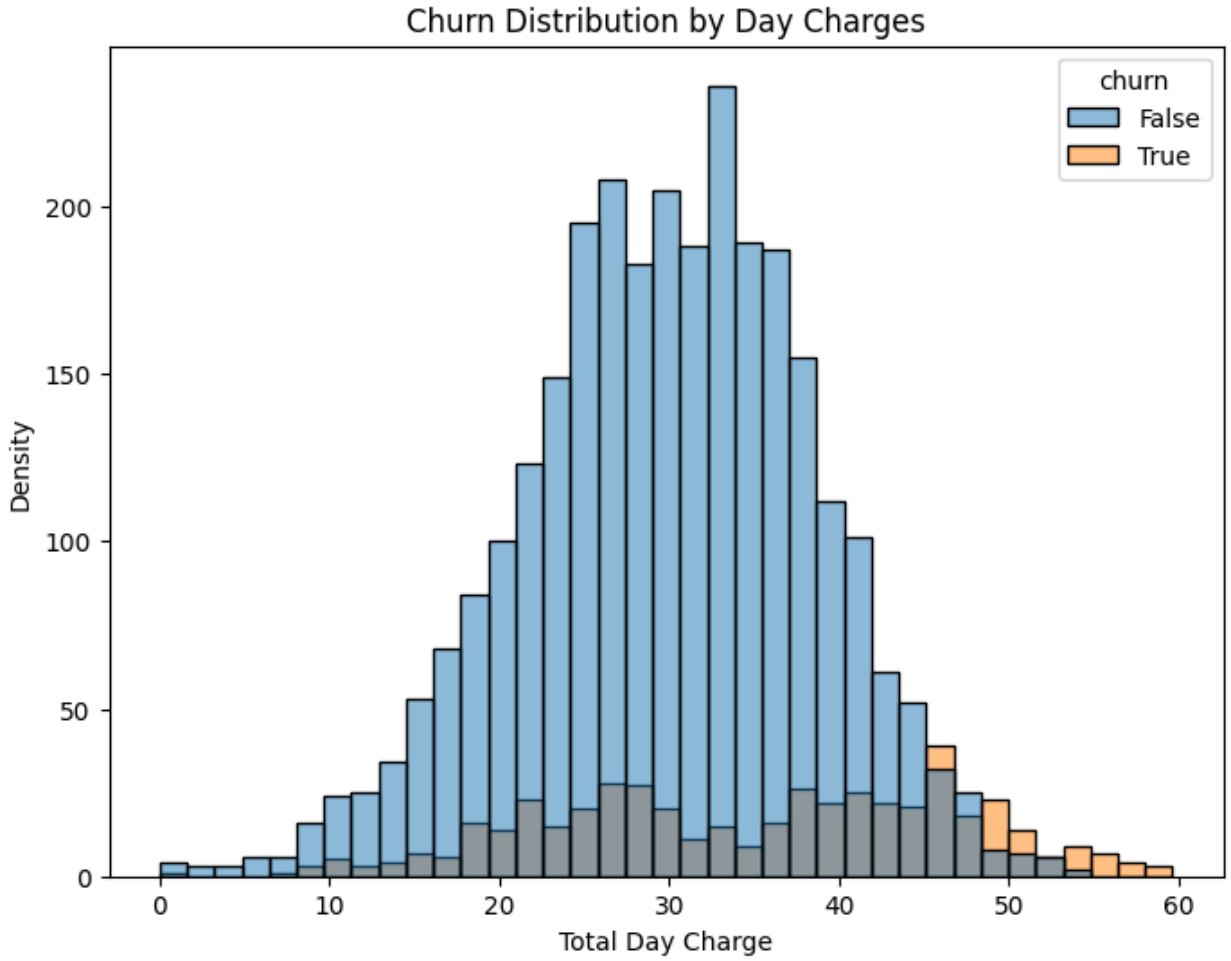
Data Analysis

- **Univariate Analysis**:
    - Distribution of churn: Approximately 85% non-churners and 15% churners (class imbalance).
    - Usage patterns of services (for example; total day minutes distribution).



The bar chart depicts the distribution of regions where customers are using telecom services. Area code 415 has the highest frequency of customers, followed by area code 510 and area code 408.

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



This graph shows that churned customers are slightly more likely to have higher total day charges compared to non-churned customers. This could indicate that customers with higher usage (and charges) might be more prone to churn.

Modelling

Models Used

1. **Baseline Model**: Decision Tree (default parameters).
    - Justification: Captures non-linear relationships with minimal preprocessing.
2. **Second Model**: Logistic Regression.
    - Justification: Offers interpretability and works well for linear relationships.
3. **Third Model**: Decision Tree with hyperparameter tuning.
    - Justification: Refines the baseline model for better performance.

Metrics of Success

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

Evaluation

Three models—a baseline Decision Tree, a hyperparameter-tuned Decision Tree, and Logistic Regression—were assessed for their ability to forecast customer attrition.

Although it showed indications of overfitting, the baseline Decision Tree was able to capture churn cases with an accuracy of 100% which showed overfitting of training data. Although it struggled with non-linear interactions in the dataset, logistic regression produced more consistent predictions with an accuracy of 75% and a recall of 73%. These early models shed light on the main predictive characteristics and the structure of the dataset.

The best-performing model was the hyperparameter-tuned Decision Tree, which had an accuracy of 87%, a recall of 80%, and a ROC-AUC of 85%. The enhanced performance shows that it can mitigate overfitting problems observed in the baseline while striking a balance between recall and precision.

The best-performing model was the hyperparameter-tuned Decision Tree, which had an accuracy of 87%, a recall of 84%, and a ROC-AUC of 87.97%. The enhanced performance shows that it can mitigate overfitting problems observed in the baseline while striking a balance between recall and precision. This model is the best option for deployment to detect and proactively retain at-risk consumers since it captures non-linear interactions efficiently.

Recommendations.

- **Target High-Risk Customers:** Based on the model's findings, prioritize retention efforts for customers with a high number of customer service calls. These customers may be

facing issues with the service, and offering proactive support or incentives could help retain them.

- **Review the International Plan:** Customers with an international plan showed a higher likelihood of churning. Consider offering tailored retention packages, such as discounts, additional features, or personalized customer support, to keep these customers engaged.
- **Personalized Engagement:** Use the churn prediction model to segment customers based on their likelihood to churn. Develop targeted marketing campaigns and personalized offers to these high-risk segments, focusing on improving customer satisfaction and loyalty.
- **Improve Customer Service Experience:** Since high customer service calls correlate with churn, addressing root causes of customer dissatisfaction (e.g., service reliability, billing issues) could help reduce churn rates. Implementing better self-service options or enhancing customer support may also alleviate frequent interactions.

Limitations

Although the model is effective at forecasting customer attrition, there are a number of drawbacks to take into account. First, it's possible that not all pertinent churn-influencing elements were included in the dataset used to train the models. For instance, although they were not considered in the investigation, outside variables such as consumer sentiment, market competitiveness, and economic conditions may have an effect. Furthermore, there are more non-churners in the dataset than churners, which can make it more difficult for the model to forecast churn for smaller classes. In comparison to other sophisticated models like random forests or gradient boosting, the Decision Tree still has limits when it comes to managing more complex relationships, even though its performance was enhanced via hyperparameter tuning.

Conclusions

Key elements that impact customer churn, like customer service encounters and the use of international plans, have been effectively identified by the churn prediction project. Particularly after hyperparameter tuning, the Decision Tree model had the highest efficacy in identifying these trends and accurately forecasting churn. The business can use this model to identify clients who are at risk and target them with customized retention measures. Even though the model works well in the current situation, its accuracy will need to be maintained through continuous data collecting and model retraining. To lower churn, enhance customer retention, and boost overall happiness, the business should take proactive measures by concentrating on the factors

that have been identified as predictors of churn, such as high customer care calls and international plan usage.

Next Steps

1. Deploy the churn prediction model for operational use.
2. Collect additional data points such as external data for example market competitions, customer satisfaction surveys or usage trends.
3. Periodically retrain the model to adapt to changing customer behavior.