ChurnQuest: A Deep Dive into Customer Retention

Project Report

Date: August 9, 2025

Owed By: Anjali Chauhan

1. Introduction and Problem Statement

1.1. Background

As a data analyst at a leading telecommunications company, I was tasked with addressing an alarming trend: an increasing number of customers discontinuing their services. This customer churn poses a significant threat to the company's revenue and market position. The "ChurnQuest" project was initiated to perform a deep-dive analysis of customer data to understand the root causes of this issue.

1.2. Problem Statement

The primary objective of this project is to **analyze the factors contributing to customer churn**. By identifying the key drivers and patterns associated with customers leaving the service, the company can develop and implement targeted, data-driven strategies to improve customer retention and satisfaction. This involves a two-pronged approach:

- **Exploratory Data Analysis:** To uncover initial insights and answer fundamental business questions through data visualization.
- **Predictive Modeling:** To build a robust machine learning model capable of identifying customers at a high risk of churning, allowing for proactive intervention.

2. Exploratory Data Analysis: Dashboard Insights

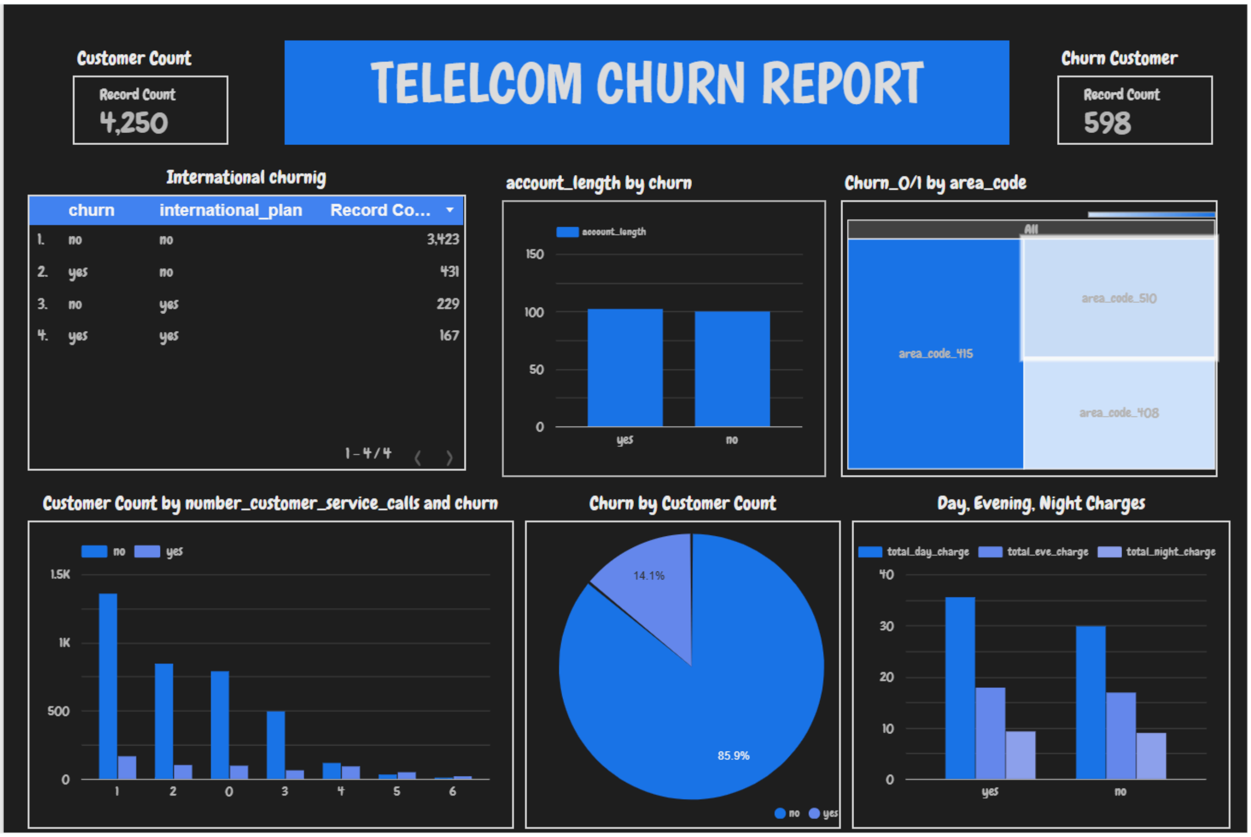
An interactive dashboard was created using **Looker Studio** to perform an initial exploration of the customer data. The dashboard provided immediate, high-level answers to several critical business questions.

2.1. Key Findings from the Looker Studio Dashboard

The following insights were gathered directly from the visualizations presented in the dashboard:

- **Overall Churn Rate:** The pie chart gives us our baseline metric: **14.1% of the total 4,250 customers have churned**, representing 598 customers lost. This highlights the significance of the problem.

- **Impact of Customer Service Calls:** The "Customer Count by number_customer_service_calls" chart is one of the most revealing. It shows a clear, strong trend: **as the number of calls to customer service increases, the proportion of customers who churn rises dramatically**. Customers with 4 or more calls have a churn rate that is nearly 50%, indicating that unresolved issues are a primary driver of customer dissatisfaction.
- **Call Charges and Churn:** The "Day, Evening, Night Charges" bar chart shows that customers who churned have noticeably **higher average charges for both day and evening calls** compared to customers who did not churn. This suggests that cost, especially for high-usage periods, is a significant factor in the decision to leave.
- **Account Tenure is Not a Key Factor:** The "account_length by churn" chart shows that the average account length for both churned and non-churned customers is almost identical. This is a crucial finding, as it tells us that **loyalty is not guaranteed with time**; both new and long-tenured customers are at risk.
- **Geographic Distribution:** The "Churn_0/1 by area_code" treemap indicates that the customer base is not evenly distributed geographically, with area_code_415 representing the largest segment of customers. While churn rates across areas were not drastically different, this allows for geographically targeted marketing if needed.



TELELCOM CHURN REPORT

Customer Count

Record Count
4,250

Churn Customer

Record Count
598

International churnig

	churn	international_plan	Record Co...
1.	no	no	3,423
2.	yes	no	431
3.	no	yes	229
4.	yes	yes	167

1 - 4 / 4

3. Technical Deep Dive: The Machine Learning Model

To move from understanding past behavior to predicting future outcomes, a machine learning model was developed. This section details the technical components of that process.

3.1. Libraries and Their Functions

The project utilized several key Python libraries:

- **Pandas:** The cornerstone for data manipulation. It was used to load the train.csv file into a DataFrame, which is essentially a structured table that allows for easy data cleaning and transformation.
- **Matplotlib & Seaborn:** These are visualization libraries used to create plots and graphs. They were instrumental in generating the feature importance chart to understand the model's decisions.
- **Scikit-learn (sklearn):** This is the core machine learning library. We used it for:
  - model_selection: To split our data into training and testing sets.
  - preprocessing: To encode and scale our data (LabelEncoder, StandardScaler).
  - ensemble: To access the RandomForestClassifier model.
  - metrics: To evaluate the model's performance (confusion_matrix,

classification_report).

3.2. Data Preprocessing Explained

Before training, the data was carefully prepared:

- **Encoding Categorical Variables (LabelEncoder):** Machine learning models work with numbers, not text. LabelEncoder was used to convert text-based features like international_plan ('yes'/'no') into numerical format (1/0).
- **Feature Scaling (StandardScaler):** Numerical features often have different scales (e.g., account_length from 1-240 vs. number_customer_service_calls from 0-9). StandardScaler rescales all numerical features to have a mean of 0 and a standard deviation of 1. This prevents features with larger ranges from unfairly dominating the model's learning process.

3.3. Model Selection: Why Random Forest?

A **Random Forest Classifier** was chosen for this task. It is an **ensemble model**, meaning it's built from many simpler models (decision trees). It was selected for several key reasons:

- **High Accuracy:** It combines the predictions of hundreds of individual decision trees to make a more robust and accurate final prediction. This "wisdom of the crowd" approach minimizes errors.
- **Handles Complexity:** It can capture complex, non-linear relationships between features that simpler models might miss.
- **Built-in Feature Importance:** It provides a direct measure of which features were most influential in its predictions, offering invaluable business insights without extra steps.

3.4. Model Training and Evaluation

The model was trained on 80% of the data and evaluated on the remaining 20% to test its performance on unseen data. The results were highly promising:

- **Overall Accuracy: 95.9%**

A more detailed look at the model's performance is provided by the **Confusion Matrix**:

	Predicted: NO Churn	Predicted: CHURN
Actual: NO Churn	716 (TN)	5 (FP)
Actual: CHURN	30 (FN)	99 (TP)

Interpretation:

- The model correctly identified **716 loyal customers** and **99 customers who would churn**.
- It only made **5 "false alarms"** (predicting a customer would churn when they wouldn't).
- Most importantly, it **missed only 30 customers** who actually churned.

This performance, particularly the high **Recall of 77%** for the churn class, indicates that the model is effective at its primary goal: identifying the majority of at-risk customers.

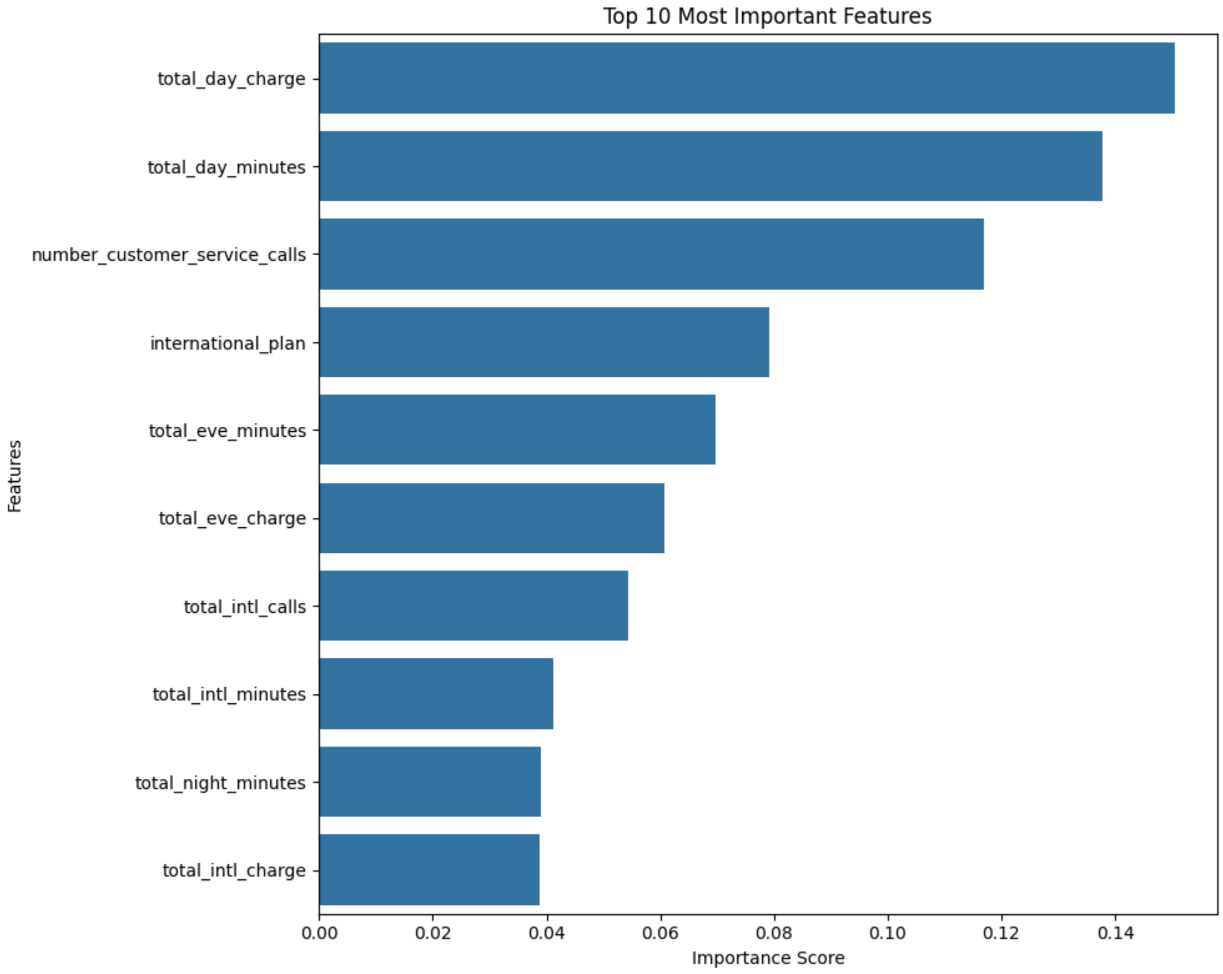
4. Key Findings and Actionable Recommendations

4.1. The Drivers of Churn: Feature Importance

The Random Forest model's built-in feature importance capability gives us a clear, quantitative ranking of the factors that most influence a customer's decision to churn. The chart below, generated directly from the model's analysis, visualizes the top 10 most significant predictors.

Top 5 Drivers from the Model:

1. **Total Day Charge & Minutes:** High daytime usage and its associated cost is the single biggest predictor of churn. The model relies on this information more than any other factor.
2. **Number of Customer Service Calls:** This is the second most powerful indicator, confirming the dashboard finding that customer frustration is a major driver.
3. **International Plan:** Whether a customer subscribes to the international plan is a top-tier predictor of churn.
4. **Total Evening Minutes & Charge:** Similar to daytime usage, high evening usage and cost is also a significant factor.
5. **Total International Calls & Charge:** The actual usage of international calling features also contributes heavily to the model's predictions.



4.2. Recommendations for a Retention Strategy

Based on the findings from both the Looker Studio dashboard and the machine learning model, the following specific, data-driven recommendations are proposed:

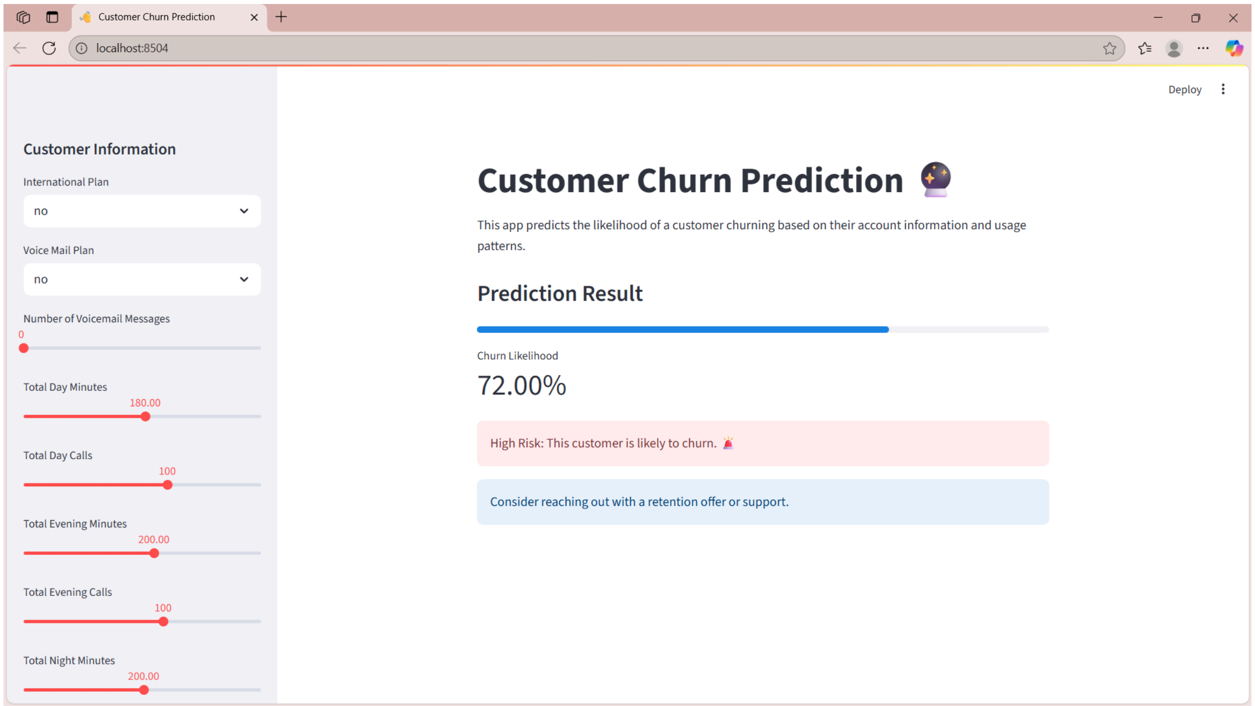
1. **Overhaul the Customer Service Process:**
   - **Finding:** The number_customer_service_calls is the second most important feature, and the dashboard shows that customers making 4 or more calls are at extremely high risk of churning.
   - **Action:** Implement a "First-Call Resolution" initiative, heavily investing in training and empowering support agents to solve problems on the initial contact. For any customer calling more than twice about the same issue, their case should be automatically escalated to a senior retention specialist.
2. **Restructure High-Usage and International Plans:**
   - **Finding:** total_day_charge is the #1 predictor of churn, and having an international_plan is the #3 predictor. This indicates significant dissatisfaction with cost and value.

    - **Action:** Introduce tiered or unlimited data plans for heavy daytime users. For the international plan, conduct A/B testing on new pricing structures or feature bundles to find a more appealing balance of cost and benefits.
3. **Launch a Proactive, Model-Driven Outreach Campaign:**
    - **Finding:** The model can predict churn with high accuracy.
    - **Action:** Using the deployed Streamlit tool to identify customers with a high churn likelihood score (e.g., >60%). Proactively contact these customers with personalized loyalty offers, such as a discount on their next bill based on their high total_day_charge, or bonus international minutes.
4. **Investigate Regional Service and Network Quality:**
    - **Finding:** While not a top churn driver overall, the dashboard shows customer distribution across area codes like 415, 510, and 408.
    - **Action:** Initiate a targeted analysis of service quality and network performance metrics in the areas with the highest customer density or emerging churn problems. Localized network issues or a lack of competitive offers in these specific areas could be contributing factors that a national-level analysis would miss.

5. Deployment: An Interactive Churn Prediction Tool

To make the model's insights accessible and actionable for business users, it was deployed as an interactive web application using **Streamlit**. This tool bridges the gap between complex data science and practical business application.

This tool allows customer service or marketing teams to input a customer's details (e.g., their usage minutes, number of service calls) and instantly receive a **churn likelihood score**. This empowers the team to make real-time, data-driven decisions, such as deciding when to offer a retention bonus to a customer on the phone.



Customer Information

International Plan: no

Voice Mail Plan: no

Number of Voicemail Messages: 0

Total Day Minutes: 180.00

Total Day Calls: 100

Total Evening Minutes: 200.00

Total Evening Calls: 100

Total Night Minutes: 200.00

Customer Churn Prediction 🔮

This app predicts the likelihood of a customer churning based on their account information and usage patterns.

Prediction Result

Churn Likelihood

72.00%

High Risk: This customer is likely to churn. 🚨

Consider reaching out with a retention offer or support.

6. Conclusion

The "ChurnQuest" project has successfully transitioned from a high-level business problem to a deployed, data-driven solution. By combining exploratory data analysis with the predictive power of machine learning, we have not only uncovered the key reasons why customers are leaving but have also built a tool to proactively identify and save at-risk customers. The implementation of the proposed recommendations is expected to significantly improve customer satisfaction and reduce the churn rate, leading to substantial long-term revenue protection.

7. Project Links and Resources

- **Looker Studio Dashboard:** Dashboard
- **GitHub Repository:** [Link to Your Project's GitHub Repository]
- **LinkedIn Profile:** [Link to Your LinkedIn Profile]