Overview

This project aims to predict customer churn using historical telecom usage data. The objective is to develop a predictive model that accurately identifies customers likely to churn, and to uncover key factors influencing churn behaviour.

Predictive Model and Features

XGBoost (Extreme Gradient Boosting) optimized machine learning library that uses an advanced form of gradient boosting to solve supervised learning problems like classification and regression. This was employed over other predictive models such as Random Forest due to its robustness, ability to handle mixed data types, and effectiveness with imbalanced classification tasks.

Features, preprocessing and data cleaning

The model leverages the key features from the original dataset. This includes call durations, plan types, customer state, and account length. To enhance predictive power, several additional features were added:

1. **totalMinutes:** Represents the total number of call minutes across all periods, providing an indicator of overall usage.
2. **totalCalls:** Captures the total number of calls made, offering another measure of customer engagement.
3. **averageCallDuration:** Calculates the average call length, as customers with longer conversations may display distinct behavioural patterns.
4. **AvgCostPerMin:** Measures the average charge per minute, helping identify potential dissatisfaction from costly or suboptimal call plans (e.g., frequent international calls).

Non-numeric attributes were converted into numerical form through categorical encoding, and uninformative columns such as *phone number* were removed.

Evaluation

As depicted in Table 1, the XGBoost churn prediction model delivers strong and reliable performance for business application.

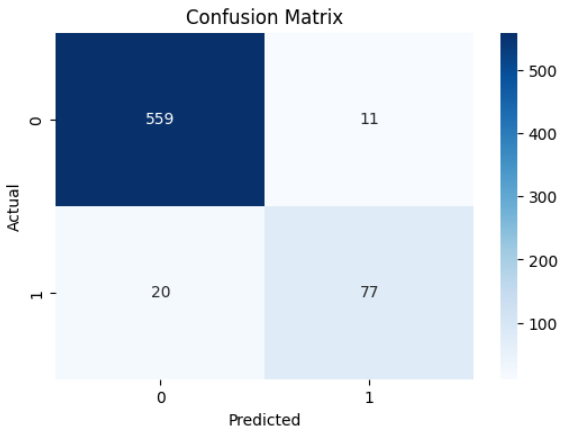
- An Accuracy of **95.35%**, the model correctly predicts the majority of customer outcomes.
- **The AUC score of 0.9046** highlights its strong ability to distinguish between customers likely to churn and those who will remain.
- An F1-score of **0.83** indicates that the model achieves a good balance between identifying churners and avoiding false alarms.
- A recall rate of **0.79** shows that it successfully detects nearly 80% of customers at risk of leaving. These results suggest that the model can be effectively used to prioritize at-risk customers for targeted retention campaigns, helping the business reduce customer attrition.

Metric	Value	Analysis
Accuracy	0.9535	Overall 95% correct predictions
AUC	0.9046	Strong ability to separate churners vs non-churners
F1 (Churn)	0.83	Balanced performance for churn prediction
Recall (Churn)	0.79	79% of actual churners detected

Table 1 Evaluation Results

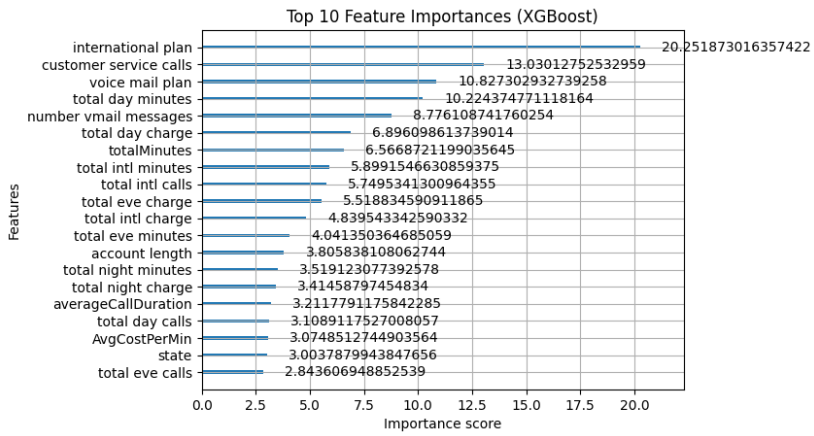
The confusion matrix demonstrates that the churn prediction model performs strongly in identifying customer loyalty and potential churn risk. Out of all test cases, **559 customers were correctly identified as non-churners**, and only **11** were incorrectly flagged as likely to churn — indicating excellent reliability in recognising customers who will stay. The model also correctly detected **77 churners**, with **20** missed cases.

These results suggest the model provides a high level of confidence for retention and loyalty programs. It enables the company to accurately target most customers at risk of leaving, while minimising false alarms that could waste retention resources. The model can meaningfully support proactive customer retention strategies and improve decision-making around marketing and service interventions.



The XGBoost model identified several key factors that most strongly influence customer churn:

1. **International Plan** emerged as the most significant predictor of churn. Customers with international plans are substantially more likely to leave, potentially due to higher costs of international usage.
2. **Customer Service Calls** ranked second, reinforcing that frequent service interactions are a strong indicator of dissatisfaction. This aligns with the insight that unresolved issues drive churn.



Insights and Predictive Indicators

1. **International Plan Customers Are the Highest-Risk Segment**
   The model identified “international plan” as the single most influential factor driving churn.

Customers subscribed to international plans appear significantly more likely to leave, suggesting potential dissatisfaction with international call pricing or perceived value.

Recommendation: This presents an actionable opportunity for the business to review and redesign international plans, improve transparency around costs, and introduce loyalty incentives or bundle offers for these customers. Targeted retention campaigns focusing on this high-risk group could substantially reduce overall churn.

2. **Frequent Customer Service Callers Indicate Unresolved Issues**
   The second most important predictor, "customer service calls," highlights that customers who contact support multiple times are strong churn indicators. This pattern suggests underlying frustration, possibly from billing disputes, technical issues, or unmet expectations.

   Recommendation: Operationally, this insight points to a need for enhanced service quality monitoring—for example, flagging accounts with repeated service interactions for proactive outreach. Implementing early-warning systems to identify and address these customers' pain points could prevent churn before it occurs.

Business Implications

This analysis enables proactive customer retention by identifying and targeting at-risk customers with tailored offers or support. It also provides valuable insights for plan redesign, such as adjusting pricing structures or service plans associated with higher churn rates. In addition, it enhances customer satisfaction tracking through the analysis of service call data, helping organizations better understand and address customer concerns.

Ethical and Governance Considerations

The model development process emphasises **fairness**, **privacy**, and **responsible use**. **Fairness** ensures that the model does not unfairly target customers from specific regions or demographics, as features such as "state" could introduce location bias. **Privacy** is protected by removing sensitive customer information, such as phone numbers, to maintain confidentiality. Finally, **responsible use** is upheld by ensuring that predictions are used to assist customer service teams rather than penalize users—for example, avoiding practices that might limit support to customers predicted to churn.

Conclusion

- The XGBoost churn model achieved **AUC = 0.90** and **F1 = 0.83**, indicating strong predictive performance.
- Key churn drivers include **customer service interactions** and frequent **international callers**.
- With proper governance, the model can support **data-driven retention** strategies while maintaining ethical standards.