MACHINE LEARNING PROJECT:
Churn Prediction Model

Course: Data Science Phase-3-Project
Project: Churn Prediction Project
Institution: Moringa School

Geoffrey Kemboi

Date : 12th February2026

Introduction

Welcome to the **Churn prediction Model**! This project, developed for the Moringa School Phase 3 Data Science curriculum. Churn prediction is a critical business objective because retaining existing customers is significantly more cost-effective than acquiring new ones

Problem Statement

Customer churn poses a major challenge for telecommunications companies, as losing customers is costly and acquiring new ones is even more expensive. The problem is to develop a predictive machine learning model that can accurately identify at-risk customers, enabling proactive retention strategies and reducing revenue loss.

Project Overview

- ✓ **Goal:** Predict customer churn using ML models
- ✓ **Business importance:** Retaining customers is cheaper than acquiring new ones
- ✓ **Business & data understanding** : Perform EDA, data clean up, and scaling
- ✓ **Approach:** Compare Logistic Regression vs. Decision Tree Classifier
- ✓ **Outcome:** Identify at-risk customers for proactive retention

Business and Data Understanding

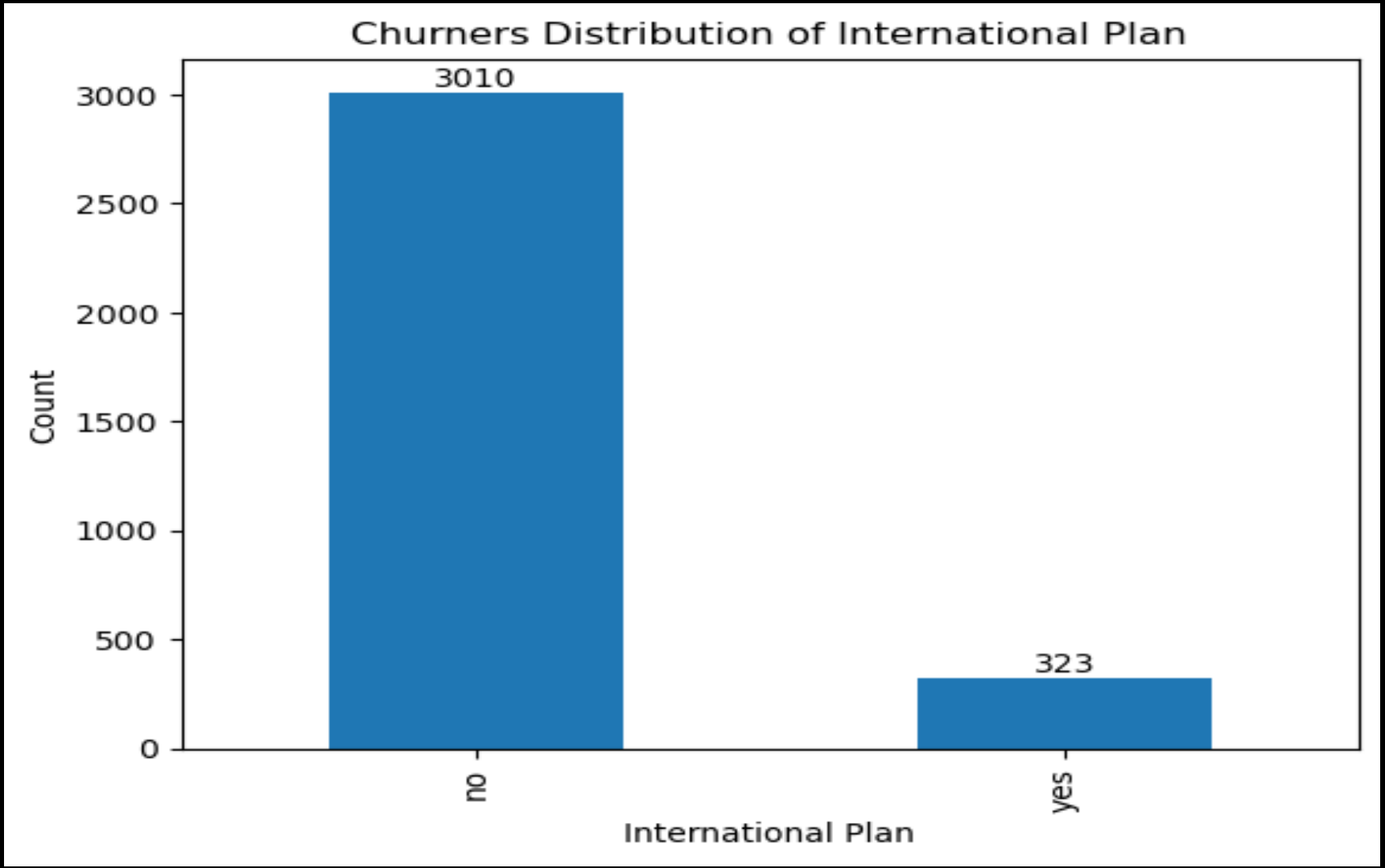
Stakeholder Audience

- Customer Success & Retention Teams
- **Need:** Flag customers likely to leave before they churn

Dataset

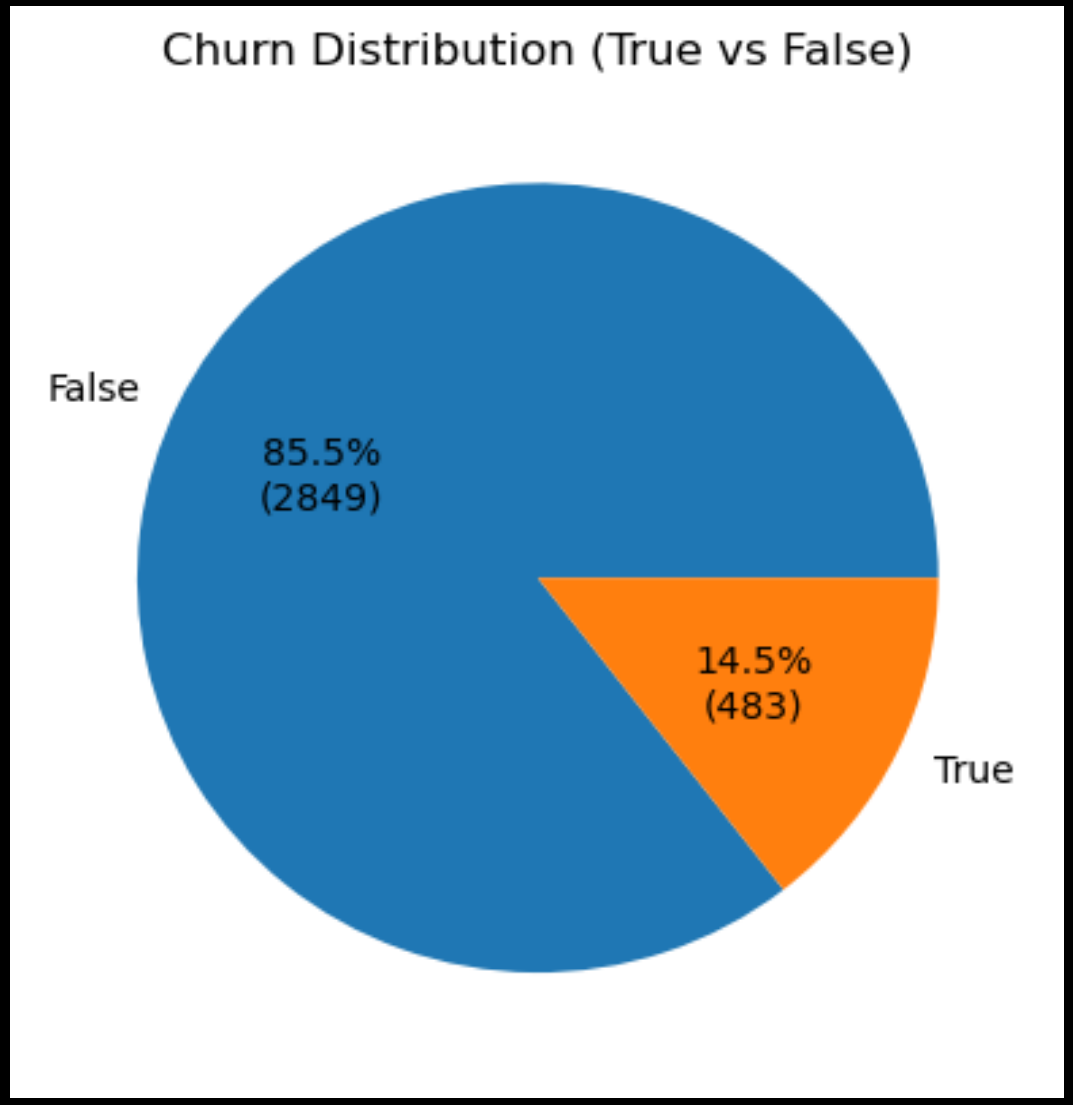
- **Source:** ChurnInTelecom.csv (3,333 records)
- **Features:** Usage patterns, account details, service interactions
- **Target:** Churn (Yes/No)
- **Challenge:** Imbalanced dataset (minority churners)

Churners based on International plan distribution



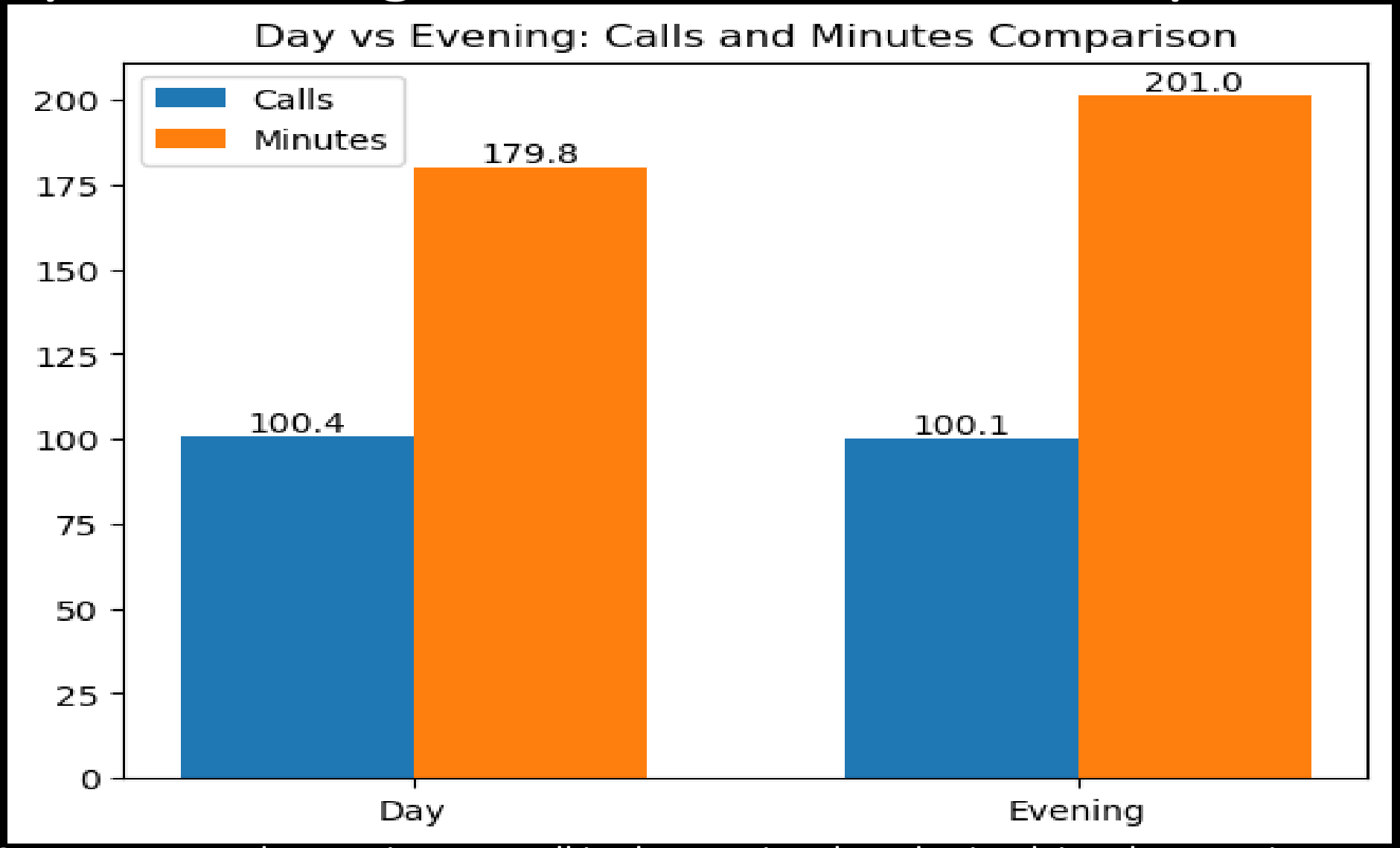
- About 3,010 churners against 323 fall into this "No International Plan" category. This is by far the larger group, indicating that most customers who churned did not subscribe to the international plan.

Churners Distribution



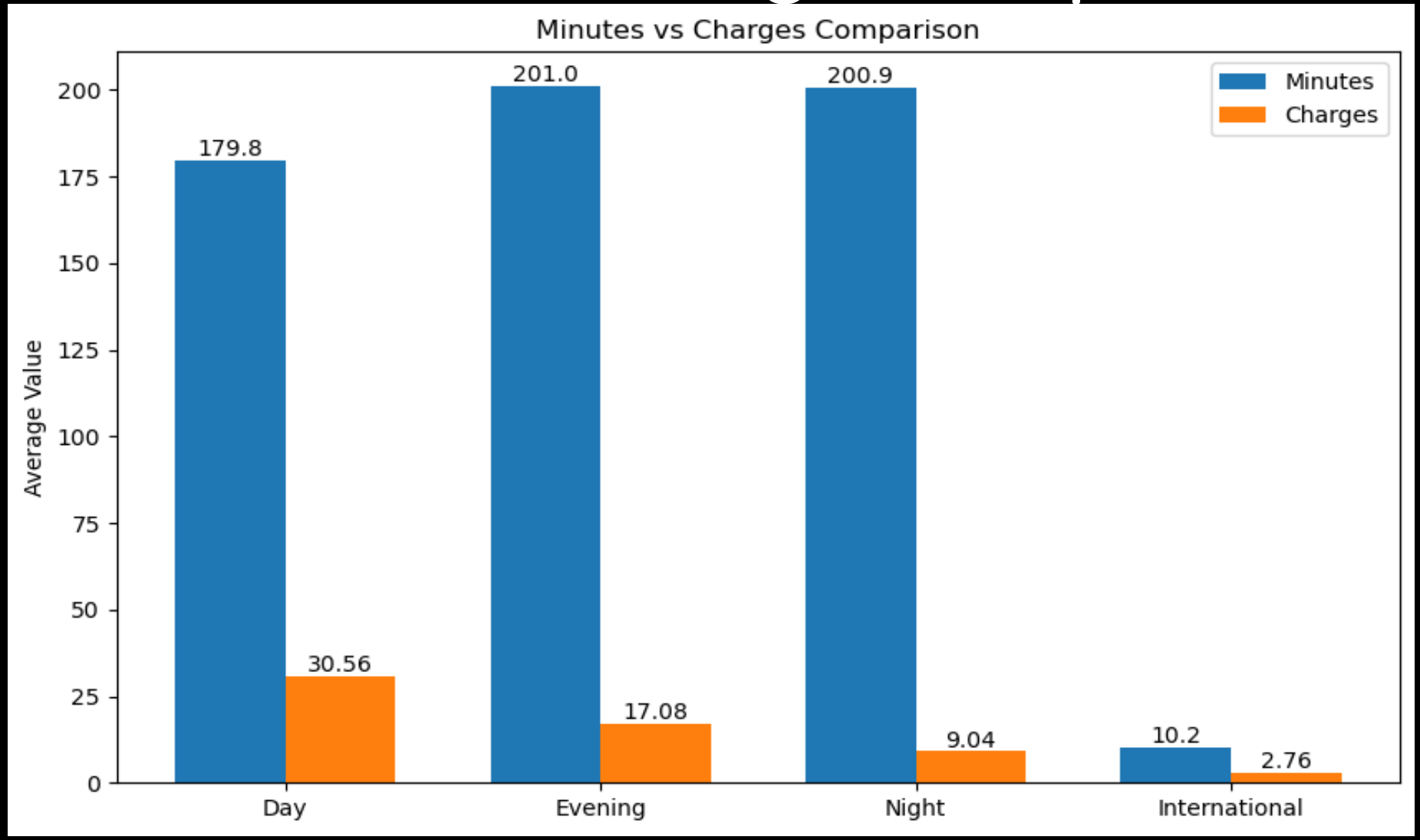
- The pie chart presentation shows that the international plan subscribers are disproportionately represented among churners.

Day vs Evening: Calls and Minutes Comparison"



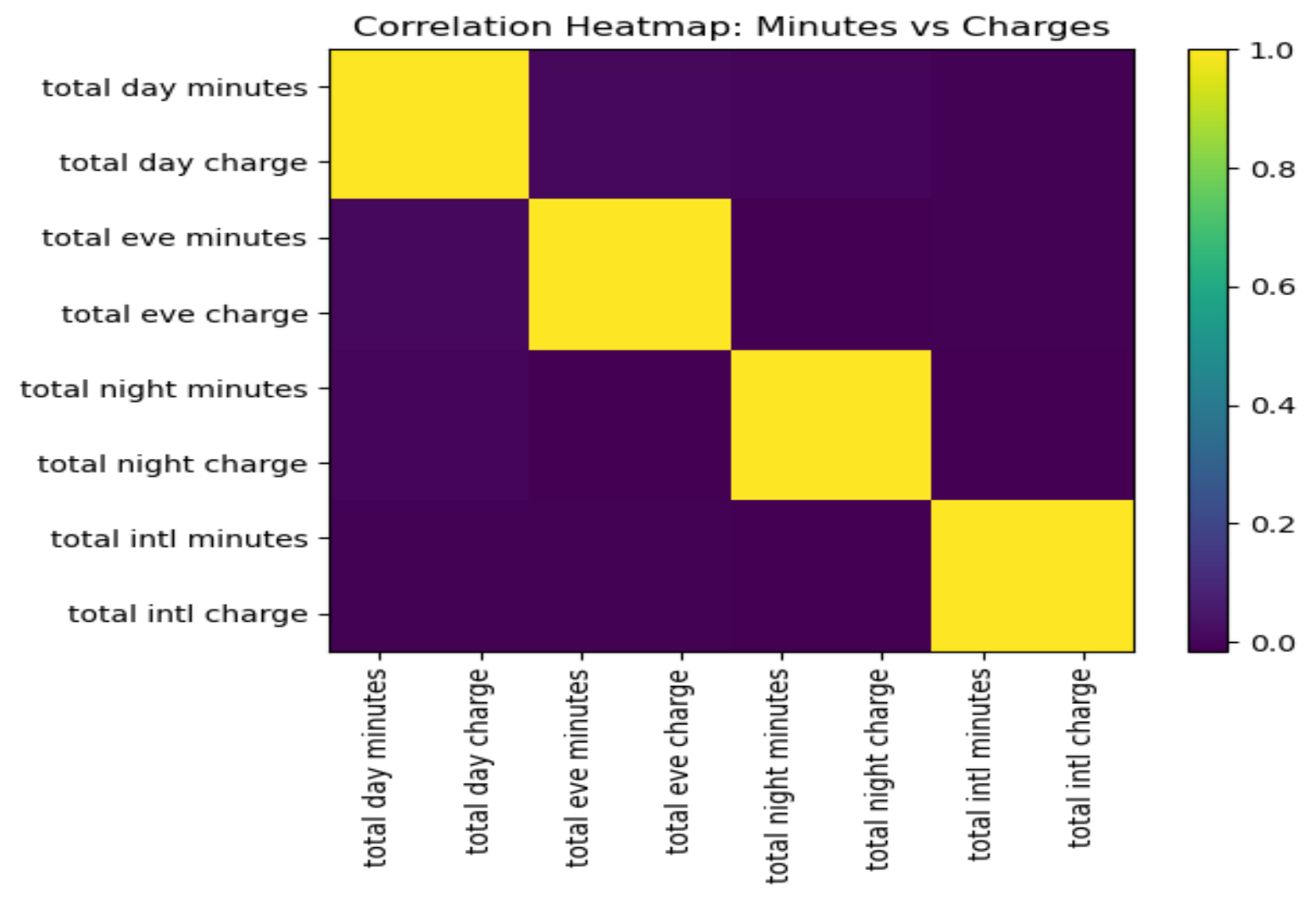
- Customers spend more time per call in the evening than day implying that evening conversations are longer on average.

Minutes vs Charges Comparison



- Customers spend the most time on evening and night calls, which are cheaper, suggesting cost-sensitive behavior. Daytime calls are shorter but more expensive, likely discouraging heavy usage While International calls are rare, but even small usage contributes to higher costs per minute.

Correlation Heatmap: Minutes vs Charges



- Charges are essentially a linear transformation of minutes, so including both in a predictive model risks multicollinearity that can distort regression results, making it better to keep minutes as the primary feature and drop charges since they add no new information.

Modelling

- Logistic Regression

- Linear, interpretable baseline

- Shows feature influence on churn odds

- Decision Tree Classifier

- Non-linear, rule-based

- Captures complex feature interactions

- Both trained on same dataset for fair comparison

Evaluation Metrics

- **Key Metrics:** Precision, Recall, F1-score
- **Performance (Both Models):**
- **Decision Tree Model**
  - Recall: ****65.90%****
  - F1 Score: ****68.76%****
  - Accuracy: ****91%****
- **Logistic Regression Model**
  - Recall: ****65.90%****
  - Accuracy: ****86%****
  - AUC: Similar to Decision Tree (~0.807)

Interpretation of Results

- **Decision Tree shows higher accuracy and F1 Score → better overall balance of precision and recall.**
- **Recall is the same for both models, indicating equal ability to identify positives.**
- **Logistic Regression has comparable AUC → similar ranking capability despite lower accuracy.**
- **Business Impact:** Enables proactive outreach to majority of churners while keeping false alarms manageable

Conclusion

- **Rationale**

- Both models outperform baseline

- Recall ensures ~2/3 of churners are identified

- **Results**

- Logistic Regression: Clear mathematical feature influence

- Decision Tree: Easier to explain with rule-based logic

Limitations

- Class imbalance still affects minority class detection

- Standard threshold may limit recall

- Models miss ~1/3 of churners

Recommendations

- Deploy Decision Tree Model for transparency and stakeholder communication

- Focus retention on high-impact features (daytime usage, multiple service calls)

- Integrate automated flagging into CRM dashboards

- Explore ensemble methods (Random Forest) to improve recall further