Problem Statement: Churn Prediction in a Complex Data Environment

A leading telecommunications company is facing a significant challenge: customer churn. The company offers a wide range of services including mobile phone plans, internet packages, streaming entertainment, and security features. However, recent analysis reveals a concerning increase in customer attrition, impacting revenue and brand reputation.

To combat churn, the company aims to implement a predictive churn model that can identify at-risk customers and proactively offer targeted retention strategies. This model will be crucial for improving customer satisfaction and ultimately driving business growth.

The data landscape presents a significant hurdle:

- Data Silos: Customer data is scattered across various systems and departments, including billing, marketing, customer service, and network operations. This fragmentation makes it difficult to obtain a complete picture of customer behavior and needs.
- Data Inconsistency: Data definitions and formats vary across systems, leading to inconsistencies and inaccuracies. For example, customer addresses may be stored differently in billing and customer service records.
- Unwanted Data: Systems often collect unnecessary data, leading to redundancy and complicating data analysis. For example, salesperson IDs or order numbers might be collected but not directly relevant for churn prediction.
- Data Security Concerns: The presence of sensitive customer data, such as social security numbers or credit card information, requires careful handling and security measures to protect customer privacy.

This complex data environment poses significant challenges for churn analysis and prediction:

- Inability to Identify True Churn Drivers: Incomplete or inaccurate data makes it difficult to identify the root causes of customer churn.
- Inaccurate Churn Predictions: Data inconsistencies and biases can lead to unreliable churn predictions, potentially resulting in ineffective retention strategies.
- Data Security Risks: Compromised sensitive customer data can damage the company's reputation and lead to legal liabilities.
- Increased Development Costs: Data cleaning and integration become time-consuming and expensive, delaying the deployment of the churn prediction model.