Project Proposal: Customer Churn Prediction in the Telecom Sector

Avani Badkul, Honey Patel, Shruti Shukla

1. **Project Overview:**

   This project focuses on using machine learning techniques to predict customer churn in the telecom industry. The primary objective is to identify behavioral and service-related factors contributing to customer attrition, allowing businesses to take proactive measures to improve retention. Churn prediction is a vital task for telecom providers, as retaining existing customers is significantly more cost-effective than acquiring new ones.

   Our work aligns with the key themes of the course: data collection, preprocessing, feature engineering, statistical analysis, and model evaluation. By applying the data science lifecycle, from data cleaning to modeling and interpretation, we aim to develop a system that supports business decisions grounded in data.

2. **Significance and problem statement:**

   Customer churn poses a critical challenge to subscription-based businesses. Telecom companies lose substantial revenue due to customers leaving unexpectedly, often without an actionable warning. Understanding the drivers behind churn, such as short contract terms, payment methods, service usage, and demographics, enables companies to design more effective engagement strategies and retention offers.

   This project also addresses common challenges in data science, such as:

   - Managing mixed-type categorical variables
   - Handling imbalanced datasets (e.g., fewer churned customers)
   - Avoiding oversimplified assumptions seen in academic datasets

   We aim to work with a realistic, noisy dataset to highlight these issues and propose practical solutions.

3. Hypothesis:

We hypothesize that customers who:

- Use month-to-month contracts
- Pay electronically
- Subscribe to fewer services (e.g., no tech support, streaming)

are significantly more likely to churn.

Conversely, customers with:

- Longer contract terms
- Multiple services (bundled phone/internet)
- Dependents or partners

are more likely to stay. We expect that modeling will confirm these associations and quantify the impact of each factor.

4. Data Plan

4.1 Dataset:

We will use the Telco Customer Churn Dataset, sourced from IBM's sample data collection. It includes:

- 7,043 customer records
- 21 attributes, including demographics, service subscriptions, account features, and a binary churn indicator

This dataset provides a well-rounded view of customer behavior and is widely used for churn modeling research.

4.2 Data Storage and Processing:

The data will be stored in a local CSV file and processed using Python (with pandas for data wrangling). Preprocessing steps include:

- Handling missing values
- Encoding categorical variables (label and one-hot encoding)
- Scaling numerical features (e.g., tenure, monthly charges)

We will also engineer new features such as:

- total_services (count of subscribed services)
- A binary flag for risky churn profiles (e.g., short tenure + high monthly charges)

These transformations are essential for model performance and interpretability.

5. **Exploratory Data Analysis (EDA):**

   EDA will be used to:

   - Explore distributions of key features (e.g., contract type, internet service, tenure)
   - Compare churn vs. non-churn distributions
   - Identify outliers, trends, and multicollinearity

   This analysis will guide feature selection and provide insight into the customer segments that are most at risk.

6. **Machine Learning Models:**

   We plan to build and compare the following models:

   - Logistic Regression

     For baseline interpretability and coefficient analysis
   - Decision Tree

     For simple rule-based churn segmentation
   - Random Forest

     For improved accuracy and handling of non-linear interactions
   - XGBoost (if time permits)

     For advanced performance and robustness

   We will split the data into training/testing sets (e.g., 80/20) and address class imbalance using:

   - Oversampling (e.g., SMOTE or simple duplication)
   - Alternative metrics (precision, recall, F1-score)

7. **Evaluation Criteria:**

   Model performance will be assessed using:

   - Accuracy
   - Precision
   - Recall (priority: catching true churn cases)
   - F1 Score
   - Confusion Matrix

   We will also interpret model outputs to identify which features most strongly influence churn probability.

8. **Conclusion:**

   This project brings together real-world business problems, technical data challenges, and predictive modeling techniques. It applies data science principles in a practical way to address a common industry concern. We anticipate our findings will highlight actionable churn risk factors and demonstrate the value of machine learning in customer relationship management.