# Customer Churn Prediction for a Subscription-Based Business

## 1. How would you handle imbalanced data if churned customers are fewer than active ones?

**Solution**

When dealing with imbalanced datasets (where churned customers are significantly fewer), several techniques can be employed:

a) Resampling Techniques:

- Oversampling the minority class (churned customers):
  - Use SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic examples of churned customers
  - Helps balance the dataset without simply duplicating existing data points
- Under sampling the majority class (active customers):
  - Randomly remove some active customer records to balance the classes
  - Can be combined with oversampling for more robust results

b) Class Weighting:

- Adjust the machine learning algorithm to give more weight to the minority class
- Particularly effective in algorithms like logistic regression, decision trees, and random forests
- Ensures the model pays more attention to the less frequent but critical churn instances

c) Ensemble Methods:

- Use algorithms specifically designed for imbalanced data
- Random Forest with balanced class weights
- Gradient Boosting with scale_pos_weight parameter
- Enables better learning from the minority class

## 2. What features are the most important predictors of churn?

**Solution**

To identify the most important features, I would recommend:

Feature Importance Analysis:

- Correlation Analysis: Examine statistical correlation between features and churn
- Permutation Importance: Measure how much model performance drops when a feature is randomly shuffled
- SHAP (Shapley Additive explanations) Values: Provide a game-theoretic approach to explain feature contributions

Potential Key Predictors:

- Tenure: Likely a strong indicator (shorter tenure might correlate with higher churn)
- Monthly Usage Hours: Low engagement could signal potential churn
- Monthly Fee: Price sensitivity might impact subscription continuation
- Subscription Plan: Different plans might have varying churn rates
- Age: Different age groups might have different retention patterns

## 3. How would you explain the model's predictions to a non-technical business team?

**Solution**

Strategies for Making the Model Interpretable:

- Use interpretable models like Decision Trees or Logistic Regression initially
- Create visual dashboards showing:
    - Key risk factors for churn
    - Probability of churn for different customer segments
    - Most influential features in prediction

Visualization Techniques:

- Confusion Matrix: Show model's prediction accuracy
- Feature Impact Charts: Graphically represent how different features influence churn probability
- Customer Segment Risk Profiles: Break down churn risk by different customer categories

## 4. What steps would you take to deploy this model into production?

Solution

Comprehensive Deployment Strategy:

a) Model Preparation

- Finalize and validate the most performant model
- Ensure model meets business performance criteria (precision, recall)
- Create a robust preprocessing pipeline

b) Infrastructure Setup

- Cloud Platform (AWS/Azure/GCP):
    - Set up scalable model serving infrastructure
    - Implement model versioning
    - Create monitoring and logging system

c) Monitoring and Maintenance

- Implement model drift detection
- Regular retraining with new data
- A/B testing of model versions
- Create alerts for significant performance changes

d) Actionable Insights Integration

- Develop automated intervention strategies
- Create personalized retention campaigns
- Trigger proactive customer engagement based on predicted risk

Recommended Model Selection:

1. Gradient Boosting (XGBoost/LightGBM)
2. Random Forest
3. Logistic Regression with regularization

Performance Metrics to Track:

- Precision
- Recall
- F1 Score
- AUC-ROC
- Confusion Matrix