PART 1: SHORT ANSWER QUESTIONS (30 POINTS)

Use Case: Customer Churn Prediction in Subscription-Based Retail

1. Problem Definition

Problem Statement
Predict which retail customers are likely to cancel their subscription within the next 30 days using behavioral and transactional data.

Objectives

- Detect at-risk subscribers early.
- Reduce churn rate through proactive campaigns.
- Increase customer lifetime value.

Stakeholders

- Marketing department
- Customer success team

Key Performance Indicator (KPI)

- **Recall for churn class**: Important to capture as many true churners as possible for effective retention (Verbeke et al., 2014).

2. Data Collection & Preprocessing

Data Sources

- Customer purchase history, frequency, and order value.
- Customer service logs (e.g., complaints, response time).

Bias Concern
Customer support bias those who complain more may be overrepresented, while passive users may go unnoticed despite being at risk of churn (Žliobaitė, 2017).

Preprocessing Steps

1. Impute missing values in behavioral metrics.
2. Normalize numerical features (e.g., purchase frequency).
3. One-hot encode features like subscription tier and region.

3. Model Development

Model Choice: Logistic Regression
Justified due to interpretability, suitability for binary classification, and speed in production systems (Hosmer et al., 2013).

Data Splitting Strategy

- 70% training, 15% validation, 15% test
- Use stratified splitting to maintain churn proportions.

Hyperparameters to Tune

- C: Controls regularization (to avoid overfitting).
- penalty: L1 or L2 depending on whether feature selection is needed.

4. Evaluation & Deployment

Evaluation Metrics

- **Recall**: Ensures most churners are identified.
- **ROC AUC**: Measures classifier performance across all thresholds (Fawcett, 2006).

Concept Drift
Customer preferences evolve (e.g., due to competitors or pricing).
Monitoring Plan: Monthly retraining, use of PSI (Population Stability Index) for drift tracking.

Technical Challenge
Real-time scoring of thousands of customers may strain infrastructure solution: batch scoring via a cloud pipeline.

PART 2: CASE STUDY APPLICATION (40 POINTS)

Use Case: Predicting Hospital Readmission Within 30 Days

1. Problem Scope

Problem: Hospitals face penalties and resource strain from high readmission rates. The goal is to build a predictive model to identify patients at high risk of being readmitted within 30 days.

Objectives

- Reduce avoidable readmissions.
- Improve discharge planning and follow-up.

- Enhance patient outcomes.

Stakeholders

- Hospital administrators
- Clinicians and discharge nurses

2. Data Strategy

Data Sources

- Electronic Health Records (EHRs)
- Discharge destination and medication count
- Comorbidities (diabetes, hypertension)

Ethical Concerns

1. **Patient Privacy**: Sensitive data must be protected (HIPAA-compliant storage and processing).
2. **Algorithmic Fairness**: Older or disabled patients may be overrepresented among readmissions, leading to potential bias in model outcomes (Obermeyer et al., 2019).

Preprocessing Pipeline

- Drop irrelevant fields (e.g., patient ID)
- Convert blood pressure into systolic/diastolic features
- Encode binary and categorical variables (e.g., gender, discharge destination)
- Normalize numerical features (e.g., age, BMI)
- Use SMOTE to address class imbalance

3. Model Development

Chosen Model: Random Forest Classifier
Robust, handles both numerical and categorical features, and performs well on imbalanced datasets when combined with class_weight = 'balanced'.

Confusion Matrix (post-SMOTE & tuning):

	Predicted No	Predicted Yes
Actual No	5123	59
Actual Yes	641	4708

Precision (Yes): 4708 / (4708 + 59) ≈ 0.99
Recall (Yes): 4708 / (4708 + 641) ≈ 0.88

4. Deployment Plan

Steps:

1. Save model with joblib or pickle
2. Deploy via API (Flask/FastAPI) connected to hospital EHR system
3. Integrate model output into clinical dashboards
4. Monitor performance regularly and retrain as needed

Compliance Strategy

- Anonymize patient records before training
- Log user access to predictions
- Secure model and data using hospital's HIPAA-compliant IT systems

5. Optimization Strategy

Method:
Apply GridSearchCV to tune key hyperparameters (max_depth, n_estimators, etc.) and cross-validation to avoid overfitting, as done in your pipeline.

PART 3: CRITICAL THINKING (20 POINTS)

1. Ethics & Bias

Bias Concern
Biased training data (e.g., underrepresentation of rural patients) may cause the model to

underpredict readmission risks in certain demographics worsening healthcare inequalities (Mehrabi et al., 2017).

Mitigation Strategy

- Analyze subgroup performance metrics (e.g., by age, gender, region)
- Use reweighting or fairness-aware algorithms to adjust model learning

2. Interpretability vs Accuracy Trade-off

Trade-off Discussion

- In healthcare, interpretability is often prioritized to gain trust and provide explainability (e.g., through feature importance).
- However, accuracy from complex models like ensembles may improve clinical utility.

Constraint Example
If computational resources are limited (e.g., in low-resource hospitals), simpler models (e.g., logistic regression) or edge deployment with smaller models may be necessary.

PART 4: REFLECTION & WORKFLOW DIAGRAM (10 POINTS)

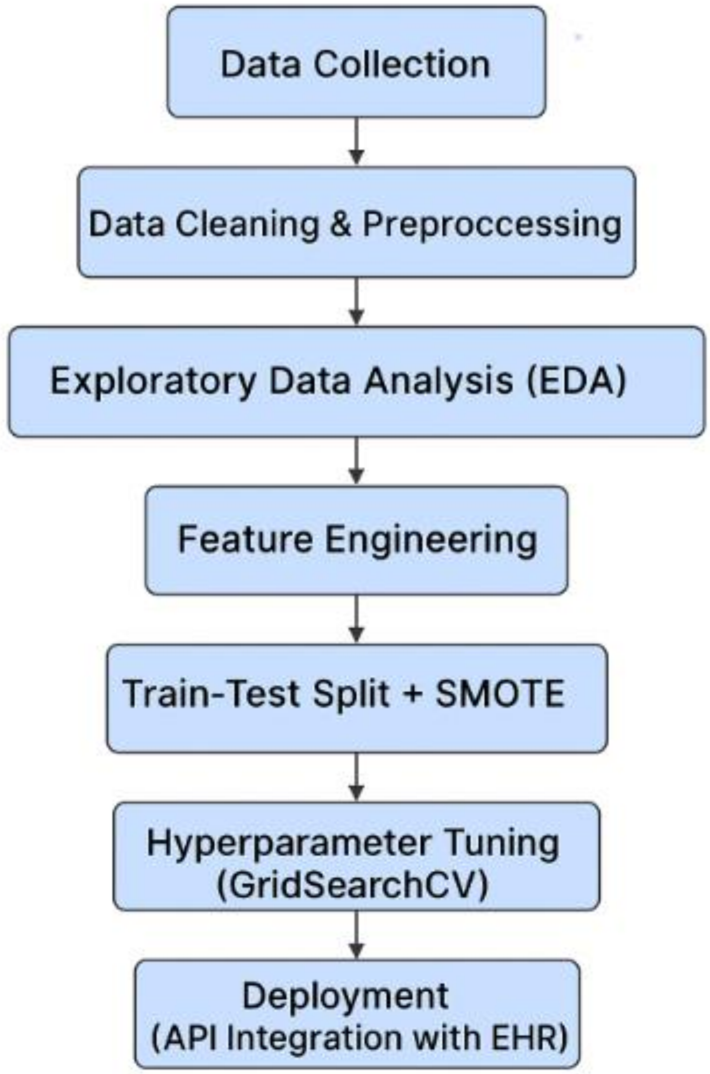
1. Reflection

Challenging Part
Handling class imbalance was the most difficult aspect. Initially, the model ignored the minority class (readmitted patients), which made evaluation metrics misleading.

Improvement Plan
With more time, I would:

- Incorporate unstructured data (e.g., discharge notes via NLP).
- Test XGBoost or LightGBM to compare performance.
- Involve clinicians in feature selection and validation.

2. AI Development Workflow Diagram



References

- Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2017). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
- Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. Applied Soft Computing, 14(2), 431–446.

- Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31(4), 1060–1089.