Student Name: Kavyasri R

Register Number: 622623205026

Institution: Shreenivasa Engineering College

Departmnnent: B.Tech. Information technology

Date of Submission: (Insert Date)

Github Repository Link.https://github.com/Kavyasri2410/Ammu

1. DATA COLLECTION AND PREPARATION:

GATHERING DATA: This involves collecting relevant data from various sources, such as:

CUSTOMER RELATIONSHIP MANAGEMENT (CRM) SYSTEMS:

Interaction history, demographics, contact details.

BILLING SYSTEMS: Subscription details, payment history, usage patterns based on subscription.

WEBSITE/APP ANALYTICS: User activity, navigation paths, time spent, features used.

CUSTOMER SERVICE INTERACTIONS: Support tickets, call logs, feedback, sentiment from text.

DATA TRANSFORMATION: Converting categorical variables into numerical formats suitable for machine learning algorithms (e.g., one-hot encoding). Scaling numerical FEATURES TO HAVE A SIMILAR RANGE
DATA TRANSFORMATION: Converting categorical variables into numerical formats suitable for machine learning algorithms (e.g., one-hot encoding). Scaling numerical features to have a similar range.
HANDLING IMBALANCED DATA: Churn datasets often have a significantly smaller proportion of churned customers compared to active customers. Techniques like oversampling the minority class, undersampling the majority class, or using synthetic data generation methods1 (e.g., SMOTE) might be necessary.

2. EXPLORATORY DATA ANALYSIS (EDA) AND PATTERN DISCOVERY:

DESCRIPTIVE STATISTICS: Calculating measures like mean, median, standard deviation to understand the distribution of features for churned and non-churned customers.
VISUALIZATION: Creating charts and graphs (e.g., histograms, scatter plots, box plots, correlation matrices) to identify initial patterns and relationships between features and churn. For example:
✨ Are customers with shorter tenures more likely to churn?
✨ Is there a correlation between the number of customer service interactions and churn?
✨ Do specific demographics exhibit higher

3. MACHINE LEARNING MODEL SELECTION
AND TRAINING:

CHOOSING APPROPRIATE ALGORITHMS:

Selecting machine learning algorithms suitable for binary classification problems (churned vs. not churned). Common choices include:

LOGISTIC REGRESSION: A linear model that estimates the probability of churn. It's interpretable and can provide insights into feature importance.

DECISION TREES: Tree-like structures that make predictions based on a series of decisions. They are easy to interpret.

RANDOM FOREST: An ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting.

GRADIENT BOOSTING MACHINES
(E.G., XGBOOST, LIGHTGBM): Powerful ensemble methods that build trees sequentially, correcting errors from previous trees. Often achieve high accuracy.

SUPPORT VECTOR MACHINES (SVM): Find a hyperplane that best separates the two classes.

NEURAL NETWORKS (DEEP LEARNING): Can learn complex patterns in large datasets, but might be less interpretable.

MODEL TRAINING: Fitting the chosen algorithm to the training data to learn the relationship between the features and the target variable (churn).
HYPERPARAMETER TUNING: Optimizing the parameters of the machine learning model using techniques like cross-validation and grid search or randomized search to achieve the best performance

4. MODEL EVALUATION AND INTERPRETATION:

CHOOSING EVALUATION METRICS: Selecting appropriate metrics to assess the model's performance, such as:
ACCURACY: The overall percentage of correct predictions.
PRECISION: The proportion of correctly identified churners out of all customers predicted as churners.
RECALL (SENSITIVITY): The proportion of actual churners that were correctly identified by the model.
F1-SCORE:The harmonic mean of precision and recall, providing a balanced measure.
AREA UNDER THE ROC CURVE (AUC): Measures the model's ability to distinguish between churners and non-churners.

MODEL INTERPRETATION: Understanding why the model is making certain predictions. Techniques for interpretation include:
FEATURE IMPORTANCE: Identifying which features have the most significant impact on the model's predictions (available in algorithms like Logistic Regression, Decision Trees, Random Forest, Gradient Boosting).
SHAP (SHAPLEY ADDITIVE EXPLANATIONS) VALUES: Providing individual explanations for each prediction by quantifying the contribution of each feature.
LIME (LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS): Explaining the predictions o f any machine learning model by approximating it locally with an interpretable model.

5. Deployment and Monitoring:

Deploying The Model: Integrating the trained model into a production system to generate predictions on new, incoming customer data. This could involve creating an API or batch processing pipeline.

MONITORING MODEL PERFORMANCE: Continuously tracking the model's accuracy and other relevant metrics over time to ensure it remains effective.

RETRAINING AND UPDATING THE MODEL: Periodically retraining the model with new data to adapt to evolvingcustomer behavior and maintain its predictive power.

ACTIONABLE INSIGHTS: The ultimate goal is to provide actionable insights to the business. This involves:
Identifying customers at high risk of churn for proactive intervention (e.g., personalized offers, improved support).
Understanding the key factors driving churn to implement strategies for improving customer retention across the board (e.g., addressing pain points, enhancing product features, improving customer service).

UNCOVERING HIDDEN PATTERNS:

The power of machine learning lies in its ability to uncover hidden patterns that traditional rule-based systems or manual analysis might miss. These patterns can involve complex interactions between multiple variables. For instance, a seemingly insignificant drop in a specific feature usage combined with a negative sentiment in recent customer feedback might be a strong predictor.