Q1.1:

Acknowledge the completeness of the provided datasets, noting the absence of a data dictionary.

Recommend creating a comprehensive data dictionary to explain the meaning and characteristics of each variable. This documentation is crucial for better understanding the features and ensuring accurate modeling.

Q1.2:

Constructing an MVP ML Code Solution:

Start by loading the datasets and performing necessary data preprocessing (e.g., handling missing values, encoding categorical variables).

Split the data into training and testing sets.

Experiment with various machine learning models suitable for classification tasks (e.g., Logistic Regression, Random Forest, Gradient Boosting) to predict churn. Use cross-validation for model evaluation and hyperparameter tuning.

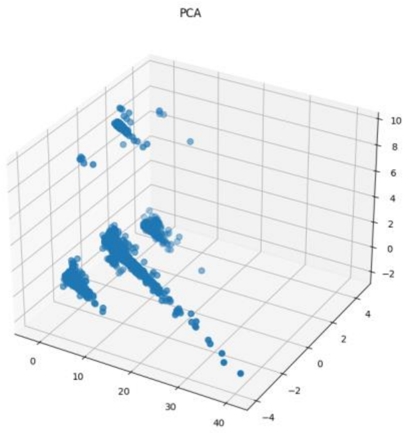
Consider feature selection techniques such as correlation analysis, chi-square tests, or using model-specific feature importance to identify influential features.

Prioritize model performance (accuracy, precision, recall, etc.) and interpretability for business insights when selecting the best-performing model.

By comparing SVM, KNN, neural network, random forest, and decision tree models, we finally found that random forest has the highest accuracy - 87.42%. I think this is because it can handle a large number of features and has a certain degree of robustness.

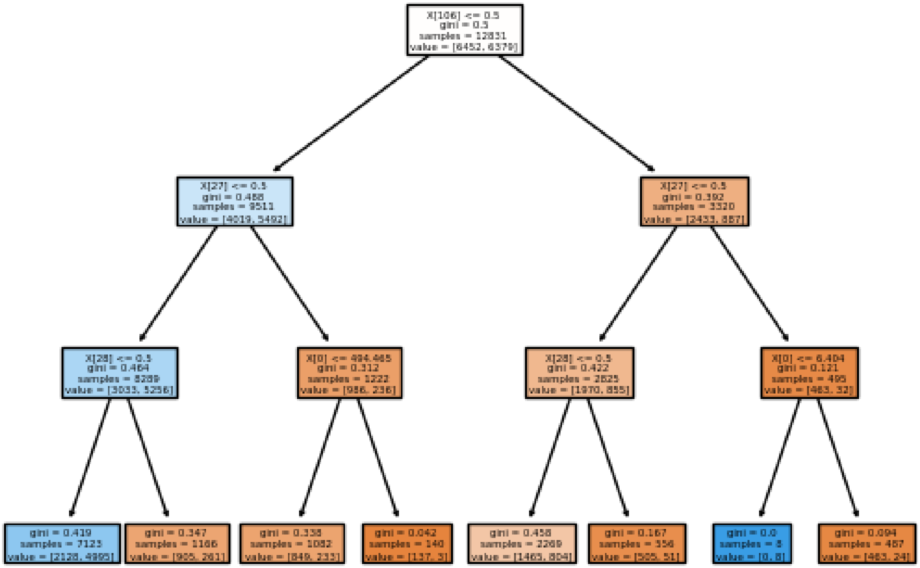
Q1.3:

I've done a Principal Component Analysis with PCA, but I don't think it's the answer we want.



Among various algorithms, because algorithms such as neural networks, SVM, and random forests are black boxes to us, decision trees and linear regression can allow us to see the proportion of each feature in decision-making. So I decided to seek help from a decision tree, which can display the desired features very intuitively.

Decision tree trained in default parameters



Features with higher importance or coefficients in the model are considered more impactful in driving churn. Analyzing the picture we can know that X[106], X[27], X[0] and X[25] have the greatest impact on customer churn.

X[106]: payment_method_cd_r_ind_current_month

X[27]: billg_prov_state_cd_ab_ind_current_month

X[0]: mrc_current_month

X[25]: srvc_prov_state_cd_sk_ind_current_month