# Customer Churn Prediction

Customer churn, also known as customer attrition or turnover, is the loss of clients or customers. It is a significant concern for businesses in various industries, such as telecommunications, internet service providers, and insurance firms, as retaining existing customers is more cost-effective than acquiring new ones.

The goal of this project is to develop a Customer Churn Prediction System that employs machine learning algorithms to predict customer churn and perform survival analysis. This system will help businesses understand the factors contributing to customer attrition and enable them to make data-driven decisions to enhance customer retention.

Data Flow Diagram :
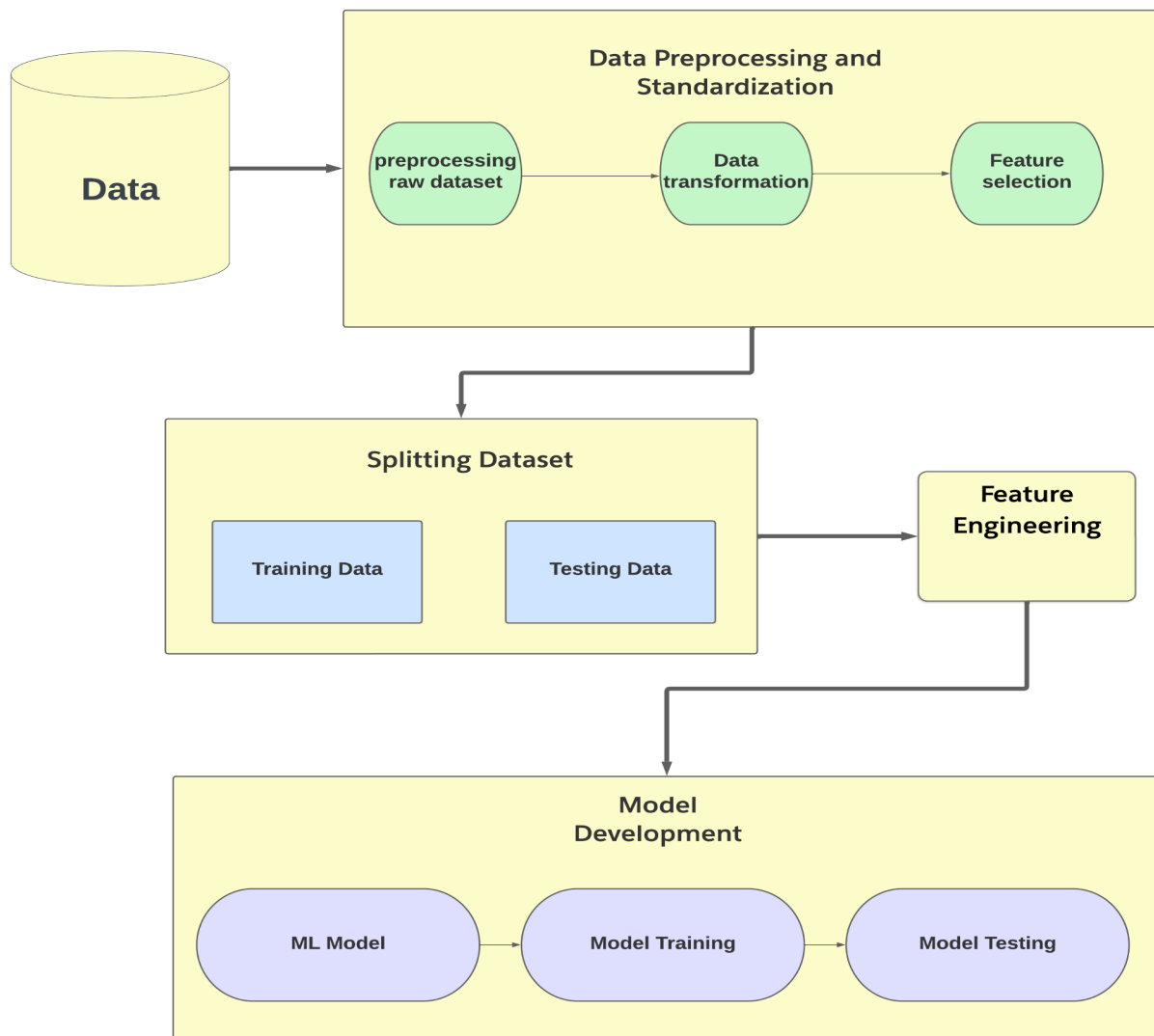
Data Collection:

Collect relevant customer data from various sources and compile it into a raw dataset for further processing and analysis.
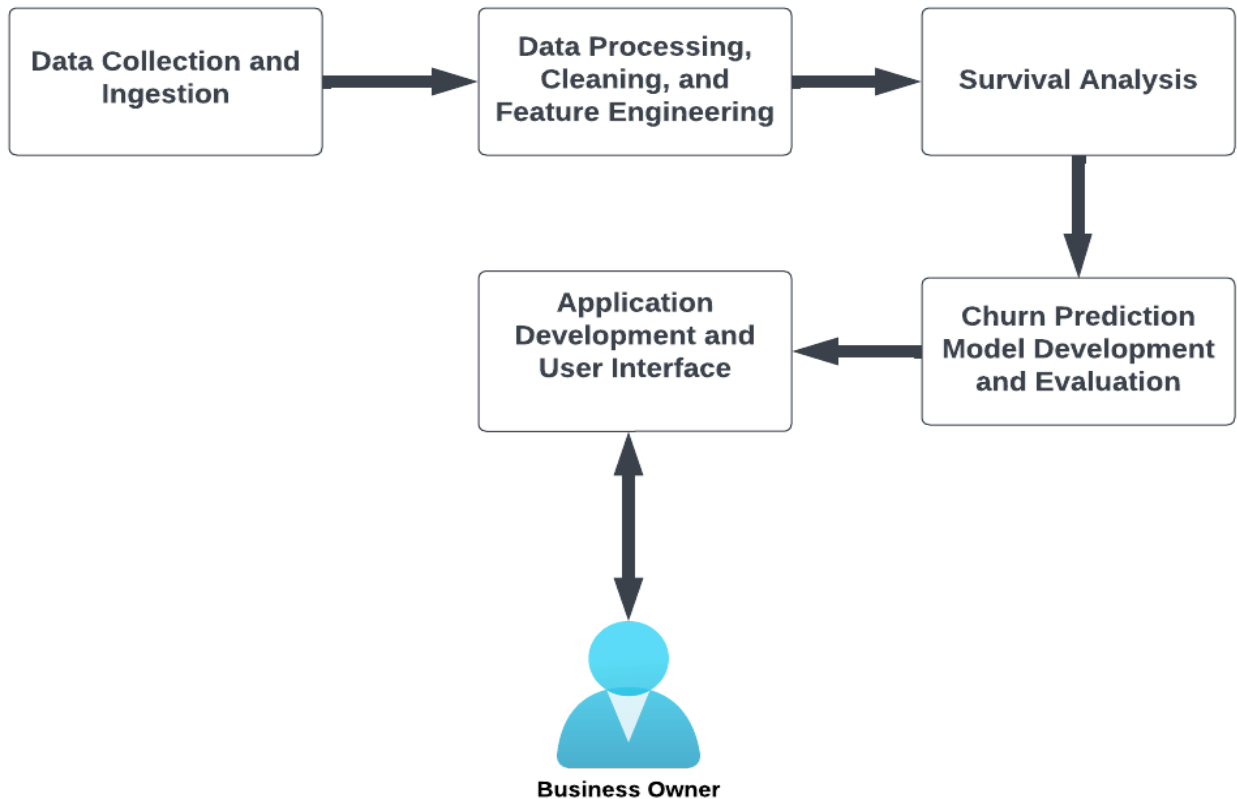
Data Preprocessing and Standardization:

- Preprocessing Raw Dataset: Clean the data by handling missing values, removing irrelevant features, and addressing inconsistencies in the data.
- Feature Engineering: Create new features based on the existing data to capture additional information that might be helpful in predicting customer churn.
- Data Transformation: Convert categorical features into numerical values using techniques like one-hot encoding, label encoding, or ordinal encoding. Normalize or standardize numerical features to ensure they are on the same scale.
- Feature Selection: Identify the most relevant features that contribute to customer churn prediction and remove redundant or less important features.

Splitting Dataset:

Divide the preprocessed dataset into two subsets: one for training the machine learning models and the other for evaluating their performance.

Model Development:

- ML Model: Select appropriate machine learning algorithms (such as logistic regression, decision trees, or support vector machines) to predict customer churn.
- Model Training: Train the selected machine learning algorithms using the training data.
- Model Testing: Test the trained models on the testing data to evaluate their performance and choose the best-performing model.

High Level Design diagram :

Data Collection and Ingestion:

- Gather relevant customer data from various sources such as demographics, purchase history, customer interactions, and product/service usage.
- Load the collected data into a suitable format (such as CSV, Excel, or a database) for further processing and analysis.

Data Processing, Cleaning, and Feature Engineering:

- Clean the data by handling missing values, removing irrelevant features, and addressing inconsistencies.
- Perform feature engineering to create new features that capture additional information helpful for predicting customer churn.
- Transform categorical features into numerical values using appropriate techniques and normalize/standardize numerical features.

Survival Analysis:

- Perform survival analysis to estimate the time until a customer churns and model the survival function for different customer segments.
- Generate Kaplan-Meier survival curves to visually represent the survival rates and compare various customer groups.

Churn Prediction Model Development and Evaluation:

- Select suitable machine learning algorithms (such as logistic regression, decision trees, or support vector machines) for churn prediction.
- Train the chosen algorithms on the preprocessed data and test their performance on a separate testing dataset.
- Evaluate the models using various metrics (e.g., accuracy, precision, recall, F1-score, AUC-ROC) and select the best-performing model.
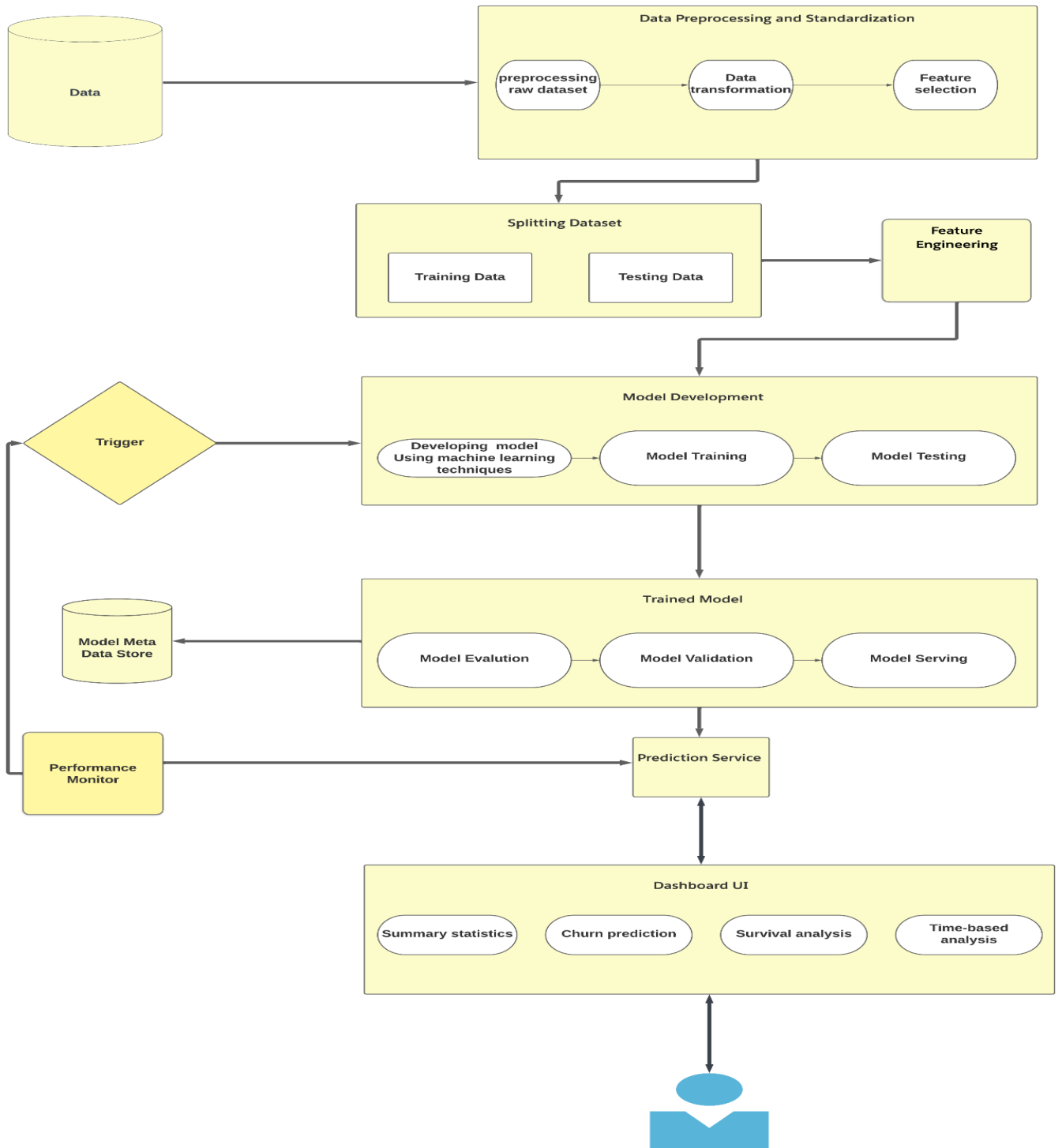
Application Development and User Interface:

- Develop a user-friendly application that allows stakeholders to interact with the churn prediction system and view insights.
- Incorporate visualizations and features such as summary statistics, churn predictions, survival analysis results, and time-based analysis.

Business Owner:

- The business owner or decision-maker will utilize the application to understand customer churn patterns and take data-driven actions to improve customer retention.
- Use the insights from the churn prediction system to identify areas of improvement, allocate resources efficiently, and inform overall business strategy.

CMPE 274

Component level design diagram:

Data

Data Preprocessing and Standardization
- preprocessing raw dataset
- Data transformation
- Feature selection

Splitting Dataset
- Training Data
- Testing Data

Feature Engineering

Trigger

Model Development
- Developing model Using machine learning techniques
- Model Training
- Model Testing

Trained Model
- Model Evaluation
- Model Validation
- Model Serving

Model Meta Data Store

Performance Monitor

Prediction Service

Dashboard UI
- Summary statistics
- Churn prediction
- Survival analysis
- Time-based analysis

Data Preprocessing and Standardization:

- Preprocessing Raw Dataset: Clean the data by handling missing values, removing irrelevant features, and addressing inconsistencies in the data.
- Feature Engineering: Create new features based on the existing data to capture additional information that might be helpful in predicting customer churn.
- Data Transformation: Convert categorical features into numerical values using techniques like one-hot encoding, label encoding, or ordinal encoding. Normalize or standardize numerical features to ensure they are on the same scale.
- Feature Selection: Identify the most relevant features that contribute to customer churn prediction and remove redundant or less important features.

Splitting Dataset:

- Training Data: A subset of the dataset used to train the machine learning models.
- Testing Data: A subset of the dataset used to evaluate the performance of the trained models.

Feature Engineering:

- TF-IDF: A text feature extraction technique that captures the importance of each term in a document relative to the entire corpus.
- Count Vectorizer: A method of converting text data into a numerical format by counting the occurrences of words in the documents.
- Ngrams:
  i. Unigrams: Individual words in a text document.
  ii. Bigrams: Pairs of consecutive words in a text document.
- Feature Sets: Collections of features that can be used to train and test the machine learning models.

Model Development:

- Developing Model Using Machine Learning Techniques: Select and implement appropriate machine learning algorithms (such as logistic regression, decision trees, or support vector machines) to predict customer churn.
- Model Training: Train the selected machine learning algorithms using the training data.
- Model Testing: Test the trained models on the testing data to evaluate their performance.
- Model Meta Data Store

Trained Model: The final, trained machine learning model that can be used to predict customer churn.

Model Evaluation:

- Model Validation: Assess the performance of the trained models using various evaluation metrics (e.g., accuracy, precision, recall, F1-score, AUC-ROC).
- Model Serving: Deploy the trained models to a production environment where they can be used to make real-time predictions.

Prediction Service:

Performance Monitor: Continuously monitor the performance of the deployed models and update them as needed to maintain accuracy and relevance.

Dashboard UI:

- Summary Statistics: Display key metrics such as the total number of customers, churned customers, churn rate, and average customer lifetime value (CLV).
- Churn Prediction: Allow users to search or filter for individual customers and view their churn probability and contributing factors.
- Survival Analysis: Show survival curves for different customer segments and compare their churn rates and median time to churn.
- Time-based Analysis: Visualize trends in churn rates over time (e.g., monthly, quarterly, or annually) and display cohort analysis results.

Team Name : Starlight
Team Members :
- **Kumar Raja Pavuluri**
- **Lokesh Vaddi**
- **Vamsidhar Reddy Menthem**
- **Thanuj Kumar Janugani**