# Lab Instructions: Customer Churn Prediction Using Logistic Regression

In this lab, you will work with a customer churn dataset to build a logistic regression model. The objective is to prepare the data, perform exploratory data analysis (EDA), select relevant features, choose an appropriate sample, build the model, and validate it. You are encouraged to use ChatGPT to guide you through each step of the process.

**Deliverables:**

One Word document that includes the lab steps. Prompts, codes, and expected inputs are provided for each step. Each step includes a short explanation of its purpose, key insights, and expected outcomes. The document also contains analysis and recommendations based on the results. Charts and screenshots are included where necessary.

## Using ChatGPT for Assistance in This Lab

Throughout this lab, ChatGPT can assist with:

* Data Preparation: Help with encoding categorical variables, handling missing values, and normalizing data.
* Exploratory Data Analysis (EDA): Suggestions for summary statistics, visualizations, and interpretations.
* Feature Selection: Identifying the most predictive variables.
* Model Building and Evaluation: Assistance in debugging code and interpreting results.

## Effective Prompting for ChatGPT

Effective prompting ensures accurate and relevant responses from ChatGPT. Poorly worded prompts can lead to vague answers, while well-structured prompts provide clear insights.

### Strategies for Writing Effective Prompts

1. Be Specific – Instead of 'How do I clean my dataset?', ask 'How do I handle missing values in numerical and categorical columns?'.
2. Provide Context – Include relevant details such as column names or sample data.
3. Iterate and Refine – If the response isn't perfect, refine your question or ask follow-ups.
4. Use Examples – Example: 'How can I encode a categorical column with values Month-to-Month, One Year, and Two Year for logistic regression?'.
5. Ask for Explanations – Instead of just code, ask 'Why is this approach recommended?'.

## Lab Steps

### Step 1: Data Preparation

* Load the dataset in Excel or Python.
* Ensure correct data formats.
* Encode categorical variables.
* Handle missing values.

### Step 2: Exploratory Data Analysis (EDA)

* Calculate summary statistics.
* Check class distribution.
* Visualize data with histograms, bar plots, and boxplots.

### Step 3: Feature Selection

* Use a correlation matrix to identify key variables.
* Drop highly correlated independent variables to reduce multicollinearity.
* Select features logically related to churn.

### Step 4: Sampling the Data

* Normalize numerical columns.
* Split into training (80%) and testing (20%) sets.

### Step 5: Building the Logistic Regression Model

* Train the model using Python’s LogisticRegression.
* Fit the model with training data.

### Step 6: Model Evaluation

* Predict churn on the test set.
* Compute Confusion Matrix: Show the number of true positives, false positives, true negatives, and false negatives.
* Visualize Confusion Matrix: Use seaborn.heatmap() for clarity.
* Calculate accuracy, precision, recall, and F1-score.
* Plot the ROC curve and compute AUC.

# Step-by-Step Guide

This guide provides step-by-step instructions to create a logistic regression model for customer churn prediction using ChatGPT. Each step includes a recommended ChatGPT prompt to assist with implementation.

## Load the Data

Load the dataset from a URL into a Pandas DataFrame and display the first few rows.

URL: https://raw.githubusercontent.com/Sid1941/data/refs/heads/main/customers\_churn.csv

ChatGPT Prompt: 'How do I load a dataset from a URL into a Pandas DataFrame and display the first few rows?'

## Data Preparation

Encode categorical variables using LabelEncoder and remove missing values from the dataset.

ChatGPT Prompt: 'How do I encode categorical variables using LabelEncoder in Python?'  
'How do I remove missing values from a Pandas DataFrame?'

## Summary Statistics

Display summary statistics for numerical columns to understand data distribution.

ChatGPT Prompt: 'How do I display summary statistics for a Pandas DataFrame?'

## Exploratory Data Analysis (EDA)

Perform visual analysis by plotting class distribution, histograms, boxplots and correlation heatmaps.

ChatGPT Prompt: 'How do I visualize class distribution and numeric variable distributions in Python?'  
'How can I create a correlation heatmap in Seaborn?'

## Feature Scaling

Standardize numerical columns using StandardScaler before training the model.

ChatGPT Prompt: 'How do I standardize numerical columns in a Pandas DataFrame using StandardScaler?'

## Splitting the Data

Split the dataset into training (80%) and testing (20%) sets for model training and evaluation.

ChatGPT Prompt: 'How do I split my dataset into training and testing sets in Python?'

## Training the Logistic Regression Model

Train a logistic regression model using Scikit-Learn on the training data.

ChatGPT Prompt: 'How do I train a logistic regression model using Scikit-Learn?'

## Making Predictions

Use the trained logistic regression model to make predictions on the test set.

ChatGPT Prompt: 'How do I make predictions using a trained logistic regression model in Python?'

## Model Evaluation

Evaluate model performance by computing the confusion matrix, accuracy, precision, recall, F1-score, and plotting the ROC curve.

ChatGPT Prompt: 'How do I evaluate a classification model in Python using a confusion matrix and ROC curve?'

## Final Notes

* Modify prompts if the response is unclear.
* Don’t hesitate to ask more questions for clarification.

## Grading Rubric (80 Points)

|  |  |  |
| --- | --- | --- |
| Criteria | Description | Points |
| Data Preparation | Loaded dataset, handled missing values, and encoded categorical variables. | 15 |
| Exploratory Data Analysis (EDA) | Conducted summary statistics, generated relevant visualizations, and interpreted results. | 15 |
| Feature Selection | Identified relevant features using correlation and logical reasoning. | 10 |
| Sampling & Model Training | Properly split data, normalized features, trained logistic regression model. | 15 |
| Model Evaluation | Assessed performance using accuracy, precision, recall, F1-score, and ROC curve. | 15 |
| Validation & Interpretation | Performed cross-validation, discussed potential overfitting, and interpreted results. | 10 |