UNIT 8

Lab Assignment Part 2: Prepare ML Models for the Real World

## Instructions

The questions below relate to the machine learning problem you would like to solve and the data set that you have chosen in the coding assignment. You will explain the type of problem you plan on solving and explain your project plan.

Except as indicated, use this document to record all your assignment work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your facilitator as part of your assignment completion. You may also have additional supporting documents that you will need to submit. Your facilitator will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used or shared outside the course, you should take care to obscure any information you feel might be of a sensitive or confidential nature.

*Begin your assignment by completing the questions below. Directions to submit your work can be found on the assignment page. Information about the grading rubric is available on any of the course assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the assignment.*

UNIT 8

# ML Problem Formulation

Answer the questions below about your machine learning problem:

## Questions:

1. List the data set you have chosen.

|  |
| --- |
| For this machine learning problem, I choose the Customer Churn dataset, which contains information about customers and whether they have churned (canceled their subscription or service) or not. |

1. What will you be predicting?

|  |
| --- |
| I will be predicting customer churn, which is a binary label indicating whether a customer is likely to churn (1) or not churn (0). |

1. Is this classification or regression problem? If this is a classification problem, is there class imbalance?

|  |
| --- |
| This is a classification problem since I am predicting a binary label (churn or not churn). |

1. Explain why this is an important problem. In other words, how would a company create value with a model that predicts this label?

|  |
| --- |
| Predicting customer churn is crucial for businesses in various industries, such as telecommunications, software as a service (SaaS), e-commerce, and more. By accurately identifying customers who are likely to churn, companies can take proactive measures to retain those customers. This could involve offering special incentives, providing personalized offers, or improving customer support, ultimately reducing customer attrition, and increasing customer loyalty. Additionally, identifying the factors that contribute to churn can help companies improve their services and customer experience. |

1. Create a project plan.

* Describe the features that you will choose.
* Choose a model (or models) that you will train.
* Explain different data preparation techniques that you may use to prepare your data for your model.
* Specify an evaluation metric that you think is appropriate for your model.
* In your plan, describe your plan to train your model, analyze its performance and then improve the model. That is, describe your model building, validation and selection plan to produce a model that generalizes well to new data.

|  |
| --- |
| * Features Selection: Select features that are relevant to predicting customer churn, such as customer demographics, usage patterns, transaction history, customer service interactions, and any other relevant data available. * Model Selection: Explore different classification models such as Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines to determine which one performs best for this specific problem. * Data Preparation Techniques: I may use techniques like handling missing values, encoding categorical variables, scaling numeric features, and handling class imbalance (if present) using methods such as oversampling or undersampling. * Evaluation Metric: Since this is a classification problem, I will use evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) to assess the model's performance. * Model Building and Validation Plan: I will split the data into training and testing sets and use cross-validation techniques to validate the model's performance, also perform hyperparameter tuning to optimize the model's parameters. To avoid overfitting, I will keep a holdout validation set to assess the model's performance on unseen data. * Model Improvement: Analyze the model's performance metrics and iterate on feature engineering, model selection, and hyperparameter tuning to improve the model's generalization and predictive power. Additionally, I will analyze feature importances to understand which features are most influential in predicting churn. * Final Model Selection: After thorough evaluation and improvement, I will select the best-performing model to deploy for production use. * Deployment: The final model will be deployed to make real-time predictions on new customer data, enabling the company to take proactive actions to prevent churn and retain valuable customers. * Monitoring: The deployed model will be monitored regularly to check for any changes in the data distribution or concept drift. If needed, the model will be retrained or fine-tuned to ensure it remains effective in predicting customer churn. |

*To submit this assignment, please refer to the instructions in the course*.