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.

|  |
| --- |
| I chose the “adult” data set, which includes Census information from 1994. |

1. What will you be predicting?

|  |
| --- |
| I will be predicting whether the income of an individual will be less than/equal to $50,000 or greater than $50,000. |

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

|  |
| --- |
| This is a classification problem because there are two cases (classes) of possible income. There is a class imbalance, as the percentage of people whose income is less than $50,000 is 75.92%, while the percentage of people with an income greater than $50,000 is 24.08%. |

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

|  |
| --- |
| This is an important problem because the information could be used by a company to target its customer base based on income levels. Also, it could be used for credit risk assessment by banks deciding whether or not a financial load can be safely given to an individual. Lastly, this information can be used by the government to identify individuals who may need financial assistance or qualify for wellfare, medicare, or medicaid programs. |

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.

|  |
| --- |
| Since the problem at hand is a binary classification problem, the models that can be used is Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier. Since there is a class imbalance, I will have to under (or over) sample the data in order to balance it out. The evaluation metrics appropriate for this model are accuracy, precision, recall, and F-1 score. The general plan will be to split the dataset into training and testing sets, evaluate the performance of each model on the testing set and then analyze those results in order to identify the best-performing model. However, if time allows for training only one model, I will most likely use Gradient Boosting Classifier, as Logistic Regression is stronger than Random Forest Classifier when working with numerical data and Gradient Boosting Classifier tends to receive higher accuracy results. Then hyperparameter tuning will be necessary using GridSearchCV. |

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