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.

| Airbnb Listings |
| --- |

1. What will you be predicting?

| If a host is a super host |
| --- |

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

| This is a binary classification problem. There could be an imbalance of more superhosts than regular hosts but that shouldn’t affect my model as much. |
| --- |

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

| For Airbnb, knowing whether a host is a super host can help you find successful Airbnb hosts and reward them/support them to have more people stay in their Airbnb and use the Airbnb platform. |
| --- |

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.

| I included 15 different columns from the 50.  host\_response\_rate float64  host\_acceptance\_rate float64  host\_is\_superhost bool  host\_listings\_count float64  host\_total\_listings\_count float64  host\_identity\_verified bool  neighbourhood\_group\_cleansed object  reviews\_per\_month float64  n\_host\_verifications int64  calculated\_host\_listings\_count int64  review\_scores\_rating float64  accommodates int64  bathrooms float64  bedrooms float64  beds float64  I will use ensemble methods and compare outputs from the various methods we’ve used. I don’t have a time constraint so models like logistic regression isn’t a priority but still could be used. I think Random forest trees might be the model I choose to go with.  As for data preparation, I need to one hot encode the neighbourhood\_group\_cleansed column but the other are numeric (with one additional boolean apart from the label column). I will also have to find the rows with most Nan values and decide if I can still harvest meaningful data without the information.  Again, I will try ensemble methods to see which model is performing the best and then tune hyperparameters before splitting off and choosing one model – If I choose to do so. |
| --- |



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