In the test set, the percentage of pass is 0.8541.

All variables include frequencies and following variables.

Before dealing with imbalanced data:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **F1 score on the test set** | **Accuracy on the test set** | **Tree plot on the training set (max depth = 5)** |
| **Absorbance data with**  **frequencies**  **only** | 0.7368 | 0.8958 |  |
| **Absorbance data with all variables** | 0.8750 | 0.9583 |  |
| **Absorbance + YAdmittance + Phase data with frequencies only** | 0.7000 | 0.8750 |  |
| **Absorbance + YAdmittance + Phase data with all variables** | 0.8750 | 0.9583 |  |

Dealing with imbalanced data (**oversampling** on the training set):

|  |  |  |  |
| --- | --- | --- | --- |
|  | **F1 score on the test set** | **Accuracy score on the test set** | **Tree plot on the training set** |
| **Absorbance data with frequencies**  **only** | 0.7778 | 0.9167 |  |
| **Absorbance data with frequencies + age** | 0.7778 | 0.9167 |  |
|  |  |  |  |

To be continued…

For next step, I will try feature selection and other resampling methods, and focus on **the Absorbance data with frequencies only.**

Q1: How important are YAdmittance, and Phase data?

Q2: Merge Absorbance, YAdmittance, and Phase VS Separate Absorbance, YAdmittance, and Phase, which one makes more sense?

From client: only focus on Absorbance frequencies.