A Brief overview of two models:

42 columns are containing missing values that play a significant role in the quality of the model. I first removed all columns with more than 40% missing values since it is almost half of the dataset (ideally, I don't think it is reliable to use half data to predict another half). I also removed categorical columns 'x39' and 'x99' since there is only 1 value in each column which is not very informative.

In the process of imputing missing values, we also need to ensure that no data leakage will occur. My original idea was to split the exercise_train.csv to train and test set; then, I can use imputing methods on two independent datasets. However, I couldn't split it with respect to all categorical features. Hence,

For logistic regression, I filled Nan values with 0 for numerical methods and replaced the missing values in categorical features with the most common values. Doing so will not leak any information in numerical features and only expose the most common values in categorical features. During the training process, I used LASSO regularization to eliminate several features and cross-validated selection of the best number of features.

I chose Catboost as the second model. Catboost has a unique way to handle categorical features and has been proved to work well in practice. It also handles NaN values automatically. Hence, there is not much to do with feature cleaning for catboost.

Comparison of two models:

Logistics regression is simple to implement and interpret. Because of the nature of the sigmoid function, $w^Tx >> 0$ or $w^Tx << 0$ implies a high probability of classification in positive or negative class, respectively. Hence, we can interpret features with large coefficients are more important than those whose coefficients are small. In addition, logistics regression doesn't make any assumptions about distributions of classes; however, the decision boundary of logistics regression is linear; hence, it tends to perform well on linear separable datasets only. Recall that there is nothing to guarantee that our data is linearly separable (as usual in practice). Since our data is high-dimensional datasets, logistics can easily overfit (that's why we used regularization techniques while training the model).

Catboost, on the other hand, can handle missing values and categorical features very well, so we don't need to do it manually. This will prevent data leakage in the pre-processing phase. However, catboost has higher complexity than logistics; hence it is more expensive in terms of computational efficiency.

Model Evaluation:

I think Catboost will perform better than logistics regression in this case. Logistics suffers in high-dimensional data, and it cannot perform well if data is not linearly separable. Second, our data has many missing values, and we must manually impute those values for logistics

regression; hence, bad imputing methods will lead to bad results. Catboost, on the other hand,

Model	AUC (test data from exercise_train.csv.)
Logistic Regression	0.7686
Catboost	0.8153

doesn't depend on how we handle missing values. We compare the AUC between two models on the test set. The result is outlined in the table below. As we can see, Catboost generates a better score than logistics regression.

Report the result:

I think the best way to report the result to a business partner is to show them the decision tree in the catboost model. It is easy to understand as it looks like a roadmap with a direction. In my experiences, the business partner would like to know how we come up with the results, so explaining the intuition of features, divide samples into groups based on features values, and the accuracy of the model will allow the business partners to have a high level of understanding of our model.

We can also tell them the false positive and false negative rates since they are intuitive in all fields.