Data Test: OperantAI

1. **The Dataset**

The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients for Taiwan from April 2005 to September 2005. The dataset has 25 variables. Objective is to predict default payment for a client based on information like sex, education, marital status, age, past payment history, and balance limit.

1. **Data Cleaning Process**

When working with real dataset, we need to take into account the fact that some data might be missing or corrupted. Therefore, the most important step is to get data cleaned and ready for analysis.

1. **Load the CSV file**:

With *read.csv* command in R. Information from the csv are stored in a dataframe *data.* There are total 30,000 rows and 25 columns.

1. **Check missing values**:

After applying required function, we checked that there are no missing values

1. **Model Fitting**

We split data into two chunks: training and testing set. The training dataset will be used to fit our model, which will be tested over testing data. In our case, we have allotted 25% of the dataset for testing purpose. Since the variable to be predicted is a binary variable 1, and 0. We use the “*Generalized Linear Model, glm*” function in “*stats*” package.

1. **Interpretation of results**

After the model was executed we analyzed the model by using “*summary*” function.

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -5.524e-01 1.380e-01 -4.004 6.24e-05 \*\*\*

LIMIT\_BAL -6.452e-07 1.812e-07 -3.560 0.000371 \*\*\*

SEX -1.144e-01 3.544e-02 -3.228 0.001245 \*\*

EDUCATION -9.834e-02 2.449e-02 -4.016 5.92e-05 \*\*\*

MARRIAGE -1.808e-01 3.640e-02 -4.968 6.78e-07 \*\*\*

AGE 4.862e-03 2.039e-03 2.385 0.017087 \*

PAY\_0 5.766e-01 2.018e-02 28.570 < 2e-16 \*\*\*

PAY\_2 7.280e-02 2.306e-02 3.156 0.001597 \*\*

PAY\_3 7.966e-02 2.571e-02 3.099 0.001945 \*\*

PAY\_4 -1.307e-02 2.908e-02 -0.450 0.653042

PAY\_5 8.026e-02 3.045e-02 2.636 0.008388 \*\*

PAY\_6 4.340e-03 2.488e-02 0.174 0.861514

BILL\_AMT1 -6.502e-06 1.350e-06 -4.815 1.47e-06 \*\*\*

BILL\_AMT2 2.956e-06 1.755e-06 1.684 0.092155 .

BILL\_AMT3 9.948e-07 1.566e-06 0.635 0.525193

BILL\_AMT4 1.034e-07 1.657e-06 0.062 0.950251

BILL\_AMT5 9.656e-07 1.833e-06 0.527 0.598264

BILL\_AMT6 4.898e-07 1.403e-06 0.349 0.727098

PAY\_AMT1 -1.476e-05 2.720e-06 -5.427 5.74e-08 \*\*\*

PAY\_AMT2 -7.229e-06 2.186e-06 -3.308 0.000940 \*\*\*

PAY\_AMT3 -4.314e-06 2.147e-06 -2.009 0.044486 \*

PAY\_AMT4 -4.051e-06 2.128e-06 -1.904 0.056897 .

PAY\_AMT5 -2.063e-06 1.976e-06 -1.044 0.296415

PAY\_AMT6 -2.723e-06 1.554e-06 -1.753 0.079672 .

The results above show that variables like BILL\_AMT1, BILL\_AMT2, BILL\_AMT3, BILL\_AMT4 are not significant in predicting the variable default.payment.next.month as they have high p-value. On the other hand, a low p-value corresponding to PAY\_0 indicates, a person who had done a payment in the last month is most likely to pay next month.

1. **Assessing the predicting ability of the model**

We assessed the predictability by plotting the ROC (Receiver Operating Characteristics) curve. A ROC curve plots results of True Positive (TP) rate and False Positive (FP) rate. If the Area Under Curve (AUC) is more than 0.5, then the model has good predictive power.