## **Credit Default Analysis Report**

## **Introduction:**

**Objective:** The goal of this report is to

* + clean the UCI default of credit cards dataset,
  + Perform regression-based exploratory analysis
  + Train and evaluate machine learning models to classify the default payment status.

**Dataset Description:**

* The dataset contains 30,000 observations and 25 variables
* No missing values were found

**Overview of Variables:**

**LIMIT\_BAL(Normalized)**: Credit Limit of clients

**SEX):** Sex of the client

**EDUCATION:** Education level of Clients

**MARRIAGE:** Marital Status of clients

**AGE:** Age of clients

**PAY\_0 to PAY\_ 6:** Past Payment status of the 6 months

**BILL\_AMT1(Normalized) to BILL \_AMT 6:** Bill amounts for the past 6 months

**PAY\_AMT1(Normalized) to PAY\_AMT6:** Payment amounts for the past 6 months

**Default payment next month(Outcome Variable)**: Indication of whether clients defaulted on their payment.

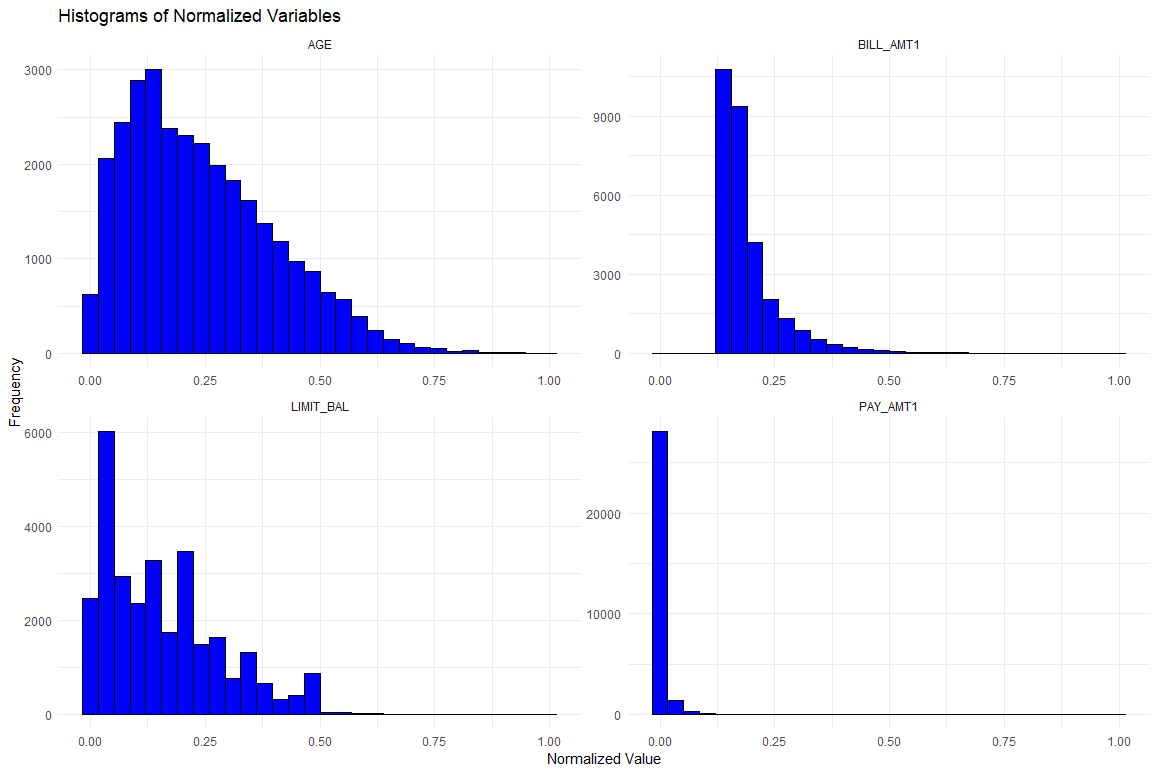
**Selected Predictors:**

Demographics: Normalized (**LIMIT\_BAL, BILL\_AMT 1, PAY\_AMT 1, AGE) variables** to improve model performance

Credit Behavior: Converted Categorical variables **(SEX, EDUCATION, MARRIAGE, default\_payment)** to factors:

These values were selected because they showed significant value in predicting the default through the regression model due to their strong correlation to default payments given the p values. There is a strong indication it affects the default payments,

**Histogram of Normalized Variables:**

****

**Exploratory Regression Analysis:**

I tested both the Logistic and probit regression models. The variables included were age, education, gender, credit limit, and payment history as the predictors. Based on the summary of the results below, Logitt regression proved to be the most accurate as it has a lower AIC score than Probit.

**Logistic Regression Model**

**Coefficients:**

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

**(Intercept) -13.1044 85.4090 -0.153 0.8781**

**LIMIT\_BAL -3.2434 0.1432 -22.656 < 2e-16 \*\*\***

**AGE 0.2063 0.1020 2.022 0.0431 \***

**SEX2 -0.1669 0.0290 -5.754 8.69e-09 \*\*\***

**EDUCATION1 11.0552 85.4077 0.129 0.8970**

**EDUCATION2 11.0631 85.4077 0.130 0.8969**

**EDUCATION3 11.0450 85.4077 0.129 0.8971**

**EDUCATION4 9.6774 85.4086 0.113 0.9098**

**EDUCATION5 9.5810 85.4081 0.112 0.9107**

**EDUCATION6 10.4813 85.4086 0.123 0.9023**

**MARRIAGE1 1.1676 0.4736 2.465 0.0137 \***

**MARRIAGE2 0.9607 0.4738 2.028 0.0426 \***

**MARRIAGE3 1.0440 0.4903 2.129 0.0332 \***

**BILL\_AMT1 1.9394 0.2616 7.412 1.24e-13 \*\*\***

**PAY\_AMT1 -22.7011 2.2686 -10.007 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for binomial family taken to be 1)**

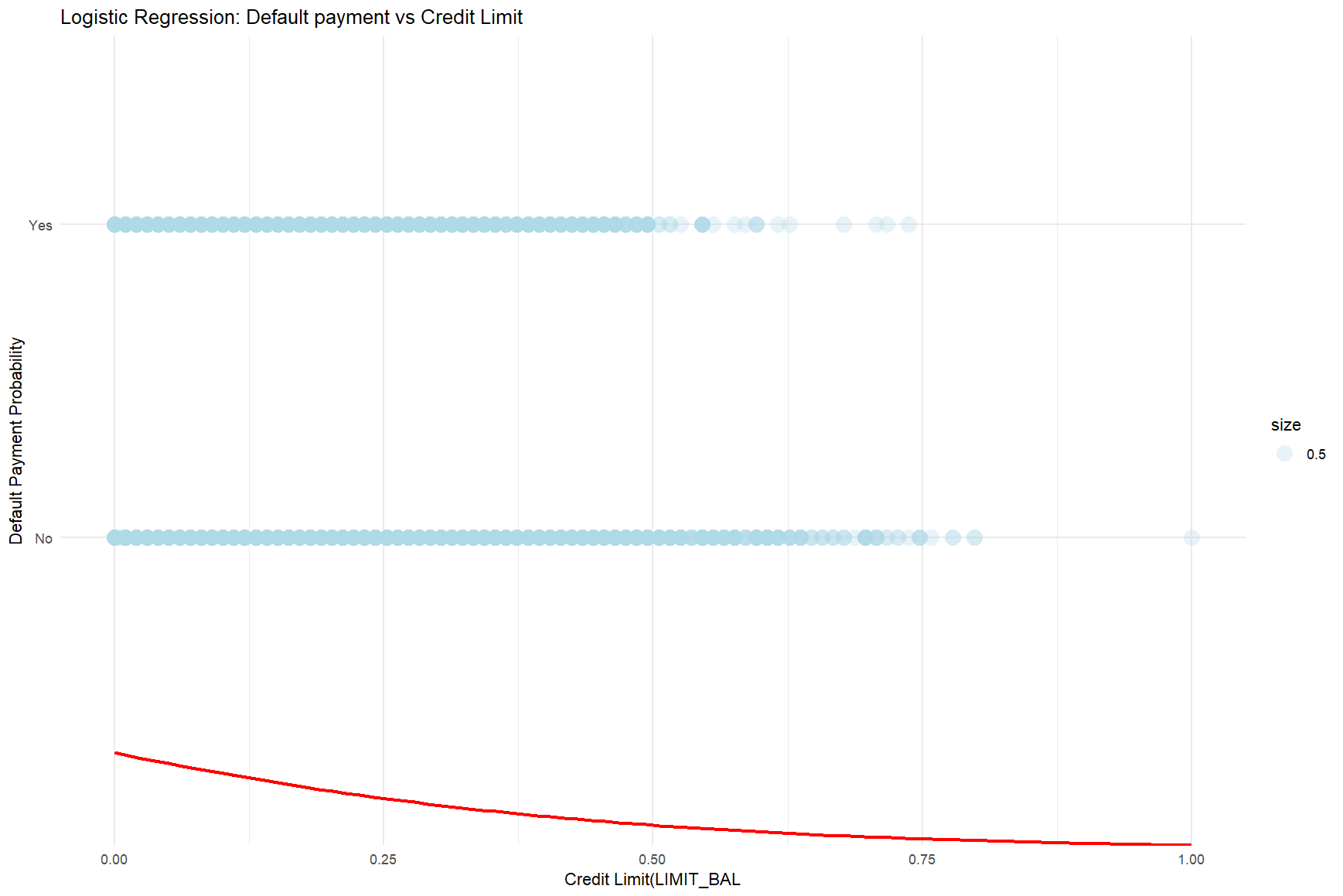
**Null deviance: 31705 on 29999 degrees of freedom**

**Residual deviance: 30541 on 29985 degrees of freedom**

**AIC: 30571**

**Number of Fisher Scoring iterations: 11**

**Scatter plot of Logit Regressions:**

****

**Probit Regression Model**

**Coefficients:**

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

**(Intercept) -4.74867 15.22552 -0.312 0.75512**

**LIMIT\_BAL -1.83893 0.07847 -23.435 < 2e-16 \*\*\***

**AGE 0.13596 0.05975 2.276 0.02287 \***

**SEX2 -0.09701 0.01685 -5.756 8.61e-09 \*\*\***

**EDUCATION1 3.55503 15.22358 0.234 0.81536**

**EDUCATION2 3.56352 15.22358 0.234 0.81492**

**EDUCATION3 3.55456 15.22359 0.233 0.81538**

**EDUCATION4 2.82176 15.22472 0.185 0.85296**

**EDUCATION5 2.76786 15.22404 0.182 0.85573**

**EDUCATION6 3.27804 15.22505 0.215 0.82953**

**MARRIAGE1 0.64732 0.24217 2.673 0.00752 \*\***

**MARRIAGE2 0.53063 0.24226 2.190 0.02850 \***

**MARRIAGE3 0.57498 0.25357 2.268 0.02336 \***

**BILL\_AMT1 0.96184 0.14355 6.700 2.08e-11 \*\*\***

**PAY\_AMT1 -8.56484 0.97525 -8.782 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for binomial family taken to be 1)**

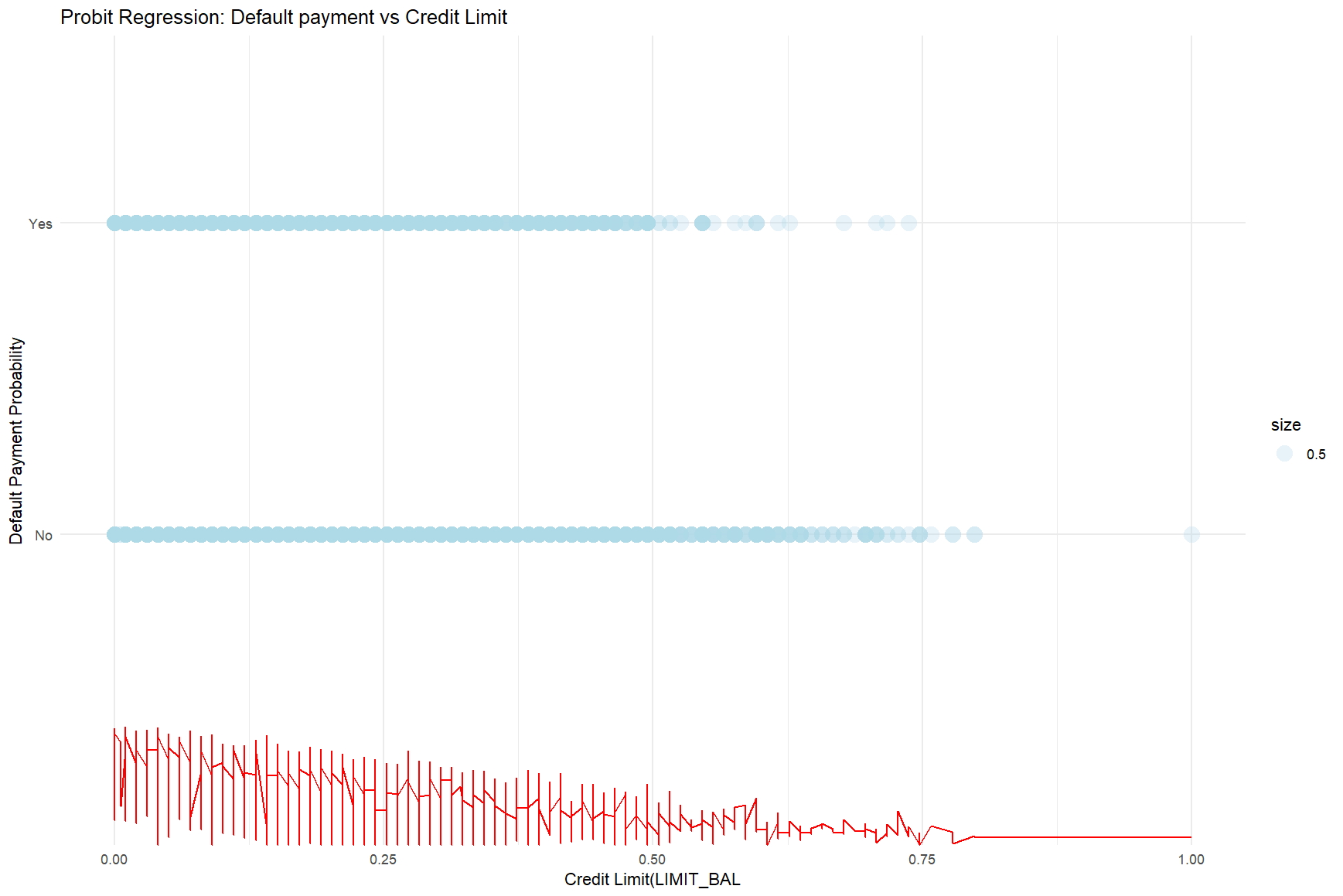
**Null deviance: 31705 on 29999 degrees of freedom**

**Residual deviance: 30585 on 29985 degrees of freedom**

**AIC: 30615**

**Number of Fisher Scoring iterations: 10**

**Scatter plot of Probit Regressions**

****

**Training a Classifier On Outcome Default**

CARET package was used for 5-fold cross-validation with a tune length of 10. SVN model provided the highest accuracy

**Training models used:**

* KNN (K Nearest Neighbors)
* SVM(Support Vector Machines)

**Model Evaluation:**

**KNN model:**

**Confusion Matrix and Statistics**

**Reference**

**Prediction No Yes**

**No 4600 1292**

**Yes 72 35**

**Accuracy : 0.7726**

**95% CI : (0.7618, 0.7832)**

**No Information Rate : 0.7788**

**P-Value [Acc > NIR] : 0.8781**

**Kappa : 0.0163**

**Mcnemar's Test P-Value : <2e-16**

**Sensitivity : 0.98459**

**Specificity : 0.02638**

**Pos Pred Value : 0.78072**

**Neg Pred Value : 0.32710**

**Prevalence : 0.77880**

**Detection Rate : 0.76679**

**Detection Prevalence : 0.98216**

**Balanced Accuracy : 0.50548**

**'Positive' Class : No**

**SVN Model:**

**Confusion Matrix and Statistics**

**Reference**

**Prediction No Yes**

**No 4672 1327**

**Yes 0 0**

**Accuracy : 0.7788**

**95% CI : (0.7681, 0.7892)**

**No Information Rate : 0.7788**

**P-Value [Acc > NIR] : 0.5074**

**Kappa : 0**

**Mcnemar's Test P-Value : <2e-16**

**Sensitivity : 1.0000**

**Specificity : 0.0000**

**Pos Pred Value : 0.7788**

**Neg Pred Value : NaN**

**Prevalence : 0.7788**

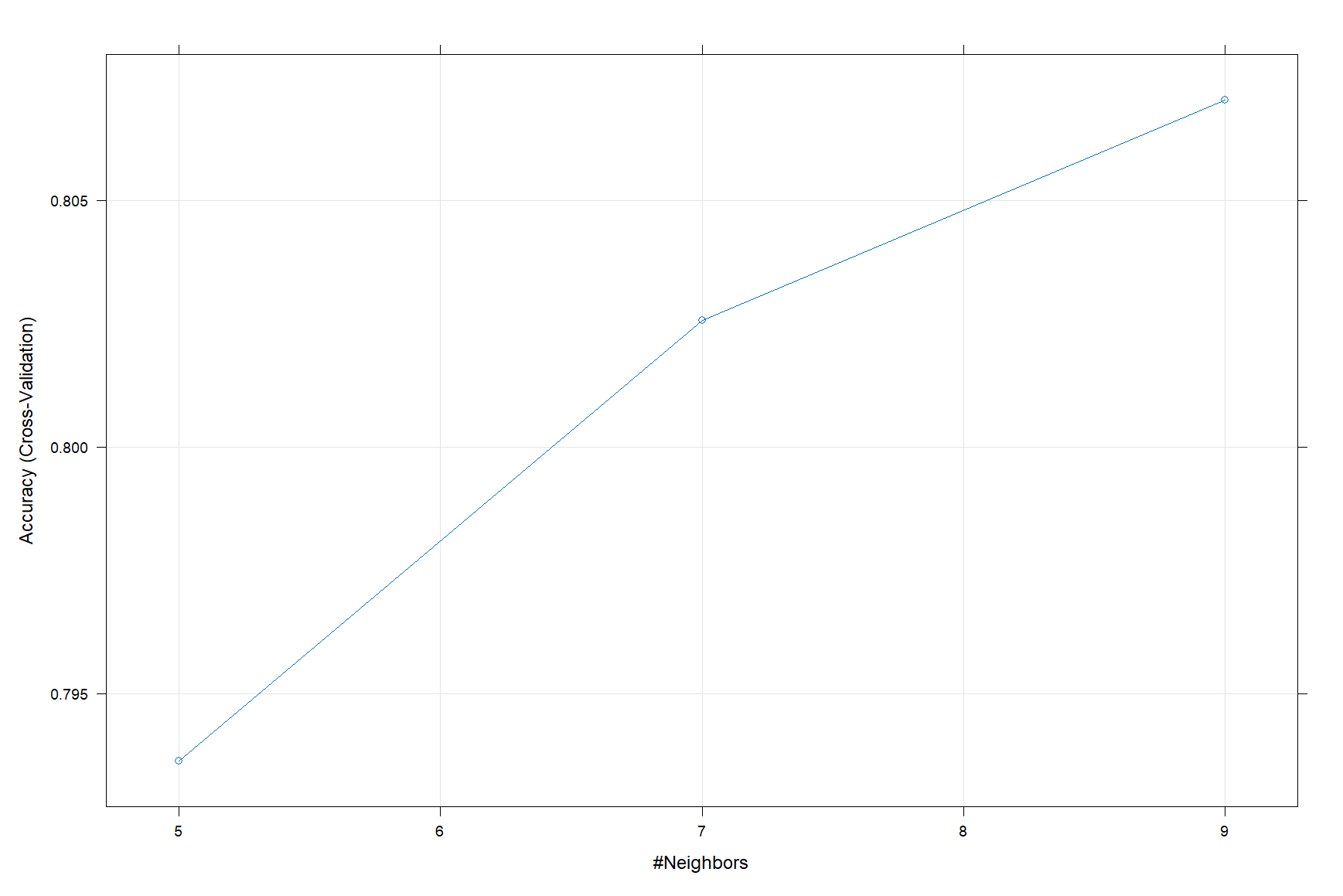
**Detection Rate : 0.7788**

**Detection Prevalence : 1.0000**

**Balanced Accuracy : 0.5000**

**'Positive' Class : No**

**Cross-validation Accuracy:**

****

**Summary:** The strongest predictors of the default payment were the Payment amount and bill amount. The strongest model choices were the Logistic regression model and the SVM model.