Revolutionizing Credit Strategies
at Martins Bank through Data Analysis

Group 3 members -

Neha Chowdary Ananthaneni - 21020366

Tanusri Makkena – 21020135

Sidharth Soibam - 21020162

Riya Dixit – 21020352

Final Project Submission for Predictive Analytics

Under the Guidance of Professor Ritanjali Panigrahi

Date of Submission: 14th September 2024

TABLE OF CONTENTS

S.No	CONTENTS	Page Number
1.	Introduction	1
2.	Data Source	3
3.	Methodology	5
4.	Results and Discussion	7
5.	Suggestions	13
6.	Conclusion	14
7.	Appendix	15

INTRODUCTION

In the wake of the 2008 financial crisis, Martins Bank, a small but ambitious institution, found itself at a crossroads. CEO Jonathan Martins believed in the power of credit, but unlike other banks, he wanted to offer credit cards in a way that managed risk without compromising accessibility.

Martins Bank Vision: *A credit card system built on data-driven decision-making to minimize defaults.*

Martins Bank had a wealth of customer data: spending patterns, income levels, and payment behaviors. Jonathan tasked his data science team with developing an advanced algorithm to assess creditworthiness with precision. This system analyzed detailed financial information and approved only those applicants who met its rigorous criteria. Credit limits were set based on a comprehensive analysis of each individual's financial situation, ensuring responsible lending.

Launching the Data-Driven Credit Card System

In 2012, Martins Bank launched its credit card product. Unlike traditional methods, the bank used real-time data analysis to evaluate each application. Customers were given personalized credit limits, tailored to their financial health, ensuring no one was overextended. This innovative approach quickly gained traction, with customers appreciating the personalized and responsible lending model.

Challenges and Evolution

As the customer base grew, Martins Bank noticed subtle shifts in payment behaviors. Some reliable customers were missing payments, posing a potential risk. In response, Jonathan's team created a predictive model that flagged early warning signs — minor changes in spending or income — allowing the bank to intervene before defaults occurred.

Seeing the need for even more accuracy, Martins Bank then invested in machine learning technology, enabling their system to learn from real-time customer behavior.

Business Problem:

Martins Bank's business problem is to **predict future credit card defaulters** using historical client data. By analyzing customers' past financial behavior, including credit limits, payment history, and demographics, the bank can identify patterns that signal a higher risk of default.

DATA SOURCE

Data retrieved from the " Default of Credit Card Clients - Predictive Models (kaggle.com) " on Kaggle.

Original study conducted by the **Department of Information Management, Chung Hua University**, Taiwan, focusing on predicting credit card defaults using client demographic and payment behavior data.

1. Variable Description:

- **ID**: A unique identifier for each customer.
- **LIMIT_BAL**: The credit limit of the customer, which is an important factor in determining the customer's purchasing power and potential risk.
- **SEX**: Gender of the customer (1 for male, 2 for female), which might be used to assess demographic trends.
- **EDUCATION**: The level of education attained by the customer, ranging from high school to postgraduate degrees.
- **MARRIAGE**: The marital status of the customer (1 for married, 2 for single, etc.).
- **AGE**: The age of the customer, which can be used for risk segmentation.
- **PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6**: Historical repayment status for the last six months (April 2005 - September 2005), with values indicating whether the customer was late in payments or paid on time. This is crucial for understanding a customer's payment behavior.
- **BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6**: The bill statements for the customer for the last six months (April 2005 - September 2005). These variables reflect the amount owed on the credit card during each month.
- **PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6**: The actual payment amounts made by the customer for each of the last six months (April 2005 - September 2005).

- **default.payment.next.month**: A categorical variable (yes or no) indicating whether the customer defaulted on their payment the following month, which is crucial for predicting credit risk.

2. Relevance of Data:

- **Demographics (SEX, EDUCATION, MARRIAGE, AGE)**: Understanding the demographics helps Martins Bank segment its customer base, tailor products, and identify trends in credit behaviour across different groups.
- **Credit Limit and Payment Data (LIMIT_BAL, PAY_0 to PAY_6, BILL_AMT1 to BILL_AMT6, PAY_AMT1 to PAY_AMT6)**: This is critical for assessing the customer's financial behaviour, managing credit risk, and predicting potential defaults.
- **Repayment Behaviour**: The historical repayment status (PAY_0 to PAY_6) is particularly important for understanding whether a customer has a habit of missing payments, which can lead to more accurate risk assessments.
- **Default Prediction (default.payment.next.month)**: This variable helps in predicting which customers are likely to default, allowing Martins Bank to adjust credit limits or initiate corrective actions.

METHODOLOGY

To predict the future credit card defaulters using historical client data and to improve credit risk management, logistic regression was employed as a suitable analytical method due to the binary nature of the dependent variable, default payment next month (0 for no, 1 for yes).

The Martins Bank dataset consists of 25 variables named as ID, LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6, default.payment.next.month. It consists of data from past transactions that have defaulted with 0 as NO and 1 as YES.

A random sample of the data is selected to split into 70% training and 30% testing using sample() to ensure randomness. The training data is used to build the model, while the testing data is for evaluation.

```
model <- glm(default.payment.next.month ~ LIMIT_BAL + SEX + EDUCATION +
MARRIAGE + AGE +
             PAY_0 + PAY_2 + PAY_3 + PAY_4 + PAY_5 + PAY_6 +
             BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT4 + BILL_AMT5 +
BILL_AMT6 +
             PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 +
PAY_AMT6,
            data = train, family = binomial)
```

The above line of code fits a logistic regression model using the glm() function. This code specifies the relationship with the independent variables (limit balance, bill amount repayment status, sex, education, etc.) and the dependent variable (default payment next month).

```
predictions <- predict(model, test, type = "response")

predicted_classes <- ifelse(predictions > 0.5, 1, 0)
```

We use the predict () function to generate the predicted probabilities for the existent clients in the test set. The dependent variable (default.payment.next.month) here is in binary form (0 = no default, 1 = default), therefore we convert these probabilities into class predictions. Here if the predicted probability is higher than 0.5 the client is predicted to default (1) and any below 0.5 is predicted to not default (0).

```
confusion_matrix <- table(test$default.payment.next.month, predicted_classes)
print(confusion_matrix)
```

The code of confusion_matrix helps create a table that allows us to evaluate the model's performance by comparing the predicted classes to the actual classes in the test set. It provides the counts for the following four categories –

a) True positives (TP) - When the model correctly predicts a default
b) True negatives (TN) - When the model correctly predicts a no-default
c) False positives (FP) - When the model incorrectly predicts a default (in reality client did not default)
d) False negatives (FN) - When the model incorrectly predicts a no-default (in reality client did default)

```
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("Accuracy:", round(accuracy, 4)))
```

The accuracy of the above model can be calculated by dividing the sum of the correctly classified instances (TP + TN) by the total number of instances in the test set. This code gives us an overall measure of how well the model performs at predicting clients as either defaulting or not defaulting.

RESULTS AND DISCUSSION

#Results of Logit Regression Model

```
Call:
glm(formula = default.payment.next.month ~ LIMIT_BAL + SEX +
    EDUCATION + MARRIAGE + AGE + PAY_0 + PAY_2 + PAY_3 + PAY_4 +
    PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT4 +
    BILL_AMT5 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
    PAY_AMT4 + PAY_AMT5 + PAY_AMT6, family = binomial, data = train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.297e+01  1.071e+02  -0.121 0.903561
LIMIT_BAL   -7.693e-07  1.878e-07  -4.097 4.19e-05 ***
SEX2        -1.236e-01  3.664e-02  -3.374 0.000741 ***
EDUCATION1   1.080e+01  1.071e+02   0.101 0.919622
EDUCATION2   1.073e+01  1.071e+02   0.100 0.920178
EDUCATION3   1.071e+01  1.071e+02   0.100 0.920329
EDUCATION4   9.773e+00  1.071e+02   0.091 0.927262
EDUCATION5   9.620e+00  1.071e+02   0.090 0.928402
EDUCATION6   9.724e+00  1.071e+02   0.091 0.927629
MARRIAGE1    1.222e+00  6.233e-01   1.960 0.050033 .
MARRIAGE2    1.035e+00  6.236e-01   1.659 0.097114 .
MARRIAGE3    9.826e-01  6.433e-01   1.527 0.126642
AGE          4.583e-03  2.232e-03   2.054 0.040005 *
PAY_0        5.943e-01  2.107e-02  28.210  < 2e-16 ***
PAY_2        6.725e-02  2.401e-02   2.801 0.005095 **
PAY_3        8.507e-02  2.676e-02   3.179 0.001478 **
PAY_4        2.733e-02  2.980e-02   0.917 0.358963
PAY_5        4.795e-02  3.202e-02   1.498 0.134201
```

```
PAY_5        4.795e-02  3.202e-02   1.498 0.134201
PAY_6       -3.116e-02  2.647e-02  -1.177 0.239099
BILL_AMT1   -4.516e-06  1.315e-06  -3.434 0.000594 ***
BILL_AMT2    1.147e-06  1.786e-06   0.643 0.520484
BILL_AMT3    2.047e-06  1.583e-06   1.293 0.195987
BILL_AMT4    1.494e-07  1.548e-06   0.096 0.923125
BILL_AMT5   -8.775e-07  1.842e-06  -0.476 0.633740
BILL_AMT6    1.207e-06  1.490e-06   0.810 0.417879
PAY_AMT1    -1.290e-05  2.786e-06  -4.632 3.62e-06 ***
PAY_AMT2    -1.266e-05  2.793e-06  -4.534 5.79e-06 ***
PAY_AMT3    -2.504e-06  2.102e-06  -1.192 0.233422
PAY_AMT4    -1.625e-06  1.959e-06  -0.830 0.406759
PAY_AMT5    -4.545e-06  2.176e-06  -2.089 0.036745 *
PAY_AMT6    -2.781e-06  1.593e-06  -1.745 0.080947 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22314  on 21039  degrees of freedom
Residual deviance: 19550  on 21009  degrees of freedom
AIC: 19612

Number of Fisher Scoring iterations: 11
```

Interpretations:

Intercept: The intercept is -12.97, with a high p-value (0.904), indicating that it is not statistically significant. This suggests that the baseline log-odds of default when all predictors are zero does not significantly differ from zero.

LIMIT_BAL: The coefficient is -7.693e-07 with a p-value of 4.19e-05, which is significant. This small negative coefficient indicates that as the credit limit increases, the probability of default decreases slightly. However, the effect size is very small.

SEX2: The coefficient is -0.1236 with a p-value of 0.000741, which is significant. This suggests that compared to the reference category (likely a different gender or category), being in the SEX2 category is associated with a decreased probability of default.

EDUCATION1 to EDUCATION6: All education coefficients are positive but not significant (p-values range from 0.919 to 0.928). This indicates that education levels do not have a significant impact on the probability of default in this model.

MARRIAGE1: The coefficient is 1.222 with a p-value of 0.050033, which is marginally significant. This suggests that being in the MARRIAGE1 category might be associated with a higher probability of default compared to the reference category.

MARRIAGE2: The coefficient is 1.035 with a p-value of 0.097114, which is marginally significant. This indicates a potential association with a higher probability of default, though the evidence is not strong.

MARRIAGE3: The coefficient is 0.9826 with a p-value of 0.126642, which is not significant. This implies that being in the MARRIAGE3 category does not significantly affect the probability of default.

AGE: The coefficient is 0.004583 with a p-value of 0.040005, which is significant. This positive coefficient suggests that older individuals have a slightly higher probability of default.

PAY_0: The coefficient is 0.5943 with a very low p-value (< 2e-16), indicating high significance. This large positive coefficient suggests that severe payment delays (PAY_0) are strongly associated with an increased probability of default.

PAY_2: The coefficient is 0.06725 with a p-value of 0.005095, which is significant. This suggests that payment delays in this period increase the probability of default.

PAY_3: The coefficient is 0.08507 with a p-value of 0.001478, which is significant. This indicates that payment delays in this period are also associated with a higher probability of default.

PAY_4: The coefficient is 0.02733 with a p-value of 0.358963, which is not significant. This implies that payment delays in this period do not significantly affect the probability of default.

PAY_5: The coefficient is 0.04795 with a p-value of 0.134201, which is not significant. This suggests that payment delays in this period do not have a strong effect on default.

PAY_6: The coefficient is -0.03116 with a p-value of 0.239099, which is not significant. This indicates that payment delays in this period do not significantly impact the probability of default.

BILL_AMT1: The coefficient is -4.516e-06 with a p-value of 0.000594, which is significant. This suggests that higher billing amounts in the first month are associated with a lower probability of default.

BILL_AMT2: The coefficient is 1.147e-06 with a p-value of 0.520484, which is not significant. This implies that billing amounts in the second month do not significantly affect the probability of default.

BILL_AMT3: The coefficient is 2.047e-06 with a p-value of 0.195987, which is not significant. This suggests that billing amounts in the third month do not have a meaningful impact on default.

BILL_AMT4: The coefficient is 1.494e-07 with a p-value of 0.923125, which is not significant. This indicates that billing amounts in the fourth month do not significantly influence default.

BILL_AMT5: The coefficient is -8.775e-07 with a p-value of 0.633740, which is not significant. This suggests that billing amounts in the fifth month do not affect the probability of default.

BILL_AMT6: The coefficient is 1.207e-06 with a p-value of 0.417879, which is not significant. This implies that billing amounts in the sixth month do not significantly impact default.

PAY_AMT1: The coefficient is -1.290e-05 with a p-value of 3.62e-06, which is highly significant. This indicates that higher payments in this period are associated with a decreased probability of default.

PAY_AMT2: The coefficient is -1.266e-05 with a p-value of 5.79e-06, which is highly significant. This suggests that higher payments in this period also reduce the probability of default.

PAY_AMT3: The coefficient is -2.504e-06 with a p-value of 0.233422, which is not significant. This indicates that payments in this period do not have a significant effect on default.

PAY_AMT4: The coefficient is -1.625e-06 with a p-value of 0.406759, which is not significant. This suggests that payments in this period do not significantly impact default.

PAY_AMT5: The coefficient is -4.545e-06 with a p-value of 0.036745, which is significant. This indicates that higher payments in this period are associated with a decreased probability of default.

PAY_AMT6: The coefficient is -2.781e-06 with a p-value of 0.080947, which is marginally significant. This suggests that higher payments in this period might be associated with a lower probability of default, though the evidence is not strong.

Null deviance - It is 22,314, representing the fit of the model with no predictors, serving as a baseline measure. The residual deviance, which is 19,550, indicates how well the model with predictors fits the data compared to the null model. A reduction in deviance from the null model suggests that the predictors improve the model's fit.

AIC (Akaike Information Criterion) - The derived AIC is 19,612. This criterion helps in model selection by balancing the goodness of fit with model complexity. A lower AIC value would indicate a preferable model that achieves a good fit without excessive complexity.

Significance - The predictors with p-values below 0.05 are considered significant. In this model, significant predictors include LIMIT_BAL, SEX2, PAY_0, PAY_2, PAY_3, PAY_AMT1, PAY_AMT2, and PAY_AMT5. These predictors significantly contribute to the model, indicating their strong relationship with the probability of default. On the other hand, variables such as EDUCATION levels and some BILL_AMT and PAY_AMT variables do not significantly affect the default probability.

Coefficient interpretation - It reflects the impact of each predictor on the log odds of default. Positive coefficients, like PAY_0 with a value of 0.5943, increase the log odds of default, suggesting that more severe payment delays lead to a higher probability of default. Negative coefficients, such as LIMIT_BAL with a value of -7.693e-07, decrease the log odds of default, implying that higher credit limits slightly reduce the default probability. The magnitude of these coefficients indicates the strength of their effect on the outcome.

Confusion matrix –

#print(confusion_matrix)

```
> print(confusion_matrix)
   predicted_classes
       0    1
  0 6790  219
  1 1470  481
```

		PREDICTED	
		0	1
ACTUAL	0	6790	219
	1	1470	481

Here,

- The actual 0 represents the clients <u>who did not default.</u>
- The actual 1 represents clients <u>who did default</u>
- The predicted 0 represents the predicted value of the clients <u>who will not default</u>
- The predicted 1 represents the predicted value of the clients <u>who will default</u>

<u>Metrics:</u>

1. True negatives - the model accurately predicted that 6790 clients who did not default
2. True positives - the model accurately predicted 481 clients who did default
3. False positives - the model incorrectly predicted 210 clients as defaulters when they did not default (Type 1 error)
4. False negatives - the model incorrectly predicted 1470 clients as non-defaulters when they defaulted (type II error)

The confusion matrix indicates that the model has a higher number of true negatives than true positives. This indicates that the model is good at predicting non-defaulters, but there is room for improvement in predicting defaulters (1470).

Accuracy:

```
> # Calculate accuracy
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
> print(paste("Accuracy:", round(accuracy, 4)))
[1] "Accuracy: 0.8115"
```

The derived 81.15% accuracy shows that the model performs well at predicting client behavior but needs to depict the complete story. Despite the high accuracy percentage, the model misses many clients who will default, as indicated by the above number of 1470 false negatives. There are errors in reflecting the correct number of defaulters.

The above logistic regression model was developed to predict/analyze the likelihood of default on Martins Bank clients' credit card payments. The model incorporates various predictors, including credit limit, demographic factors (sex, education, marriage status), age, payment history, bill statements, and payment amounts.

Below are the significant predictors-

1. **LIMIT_BAL:** The credit limit has an inversely proportionate relationship with the probability of default i.e., the higher the credit limit, the lower the default risk.
2. **SEX:** Gender is a significant predictor, notably affecting default probability.
3. **AGE:** The data suggest that older clients have a slightly lower risk of default.
4. **PAY_0 to PAY_6:** Past payment history is strongly correlated with default risk. Early payments show a significant impact on the likelihood of default.
5. **BILL_AMT1 and PAY_AMT1:** Recent billing and payment amounts also affect the risk of default. Higher recent payments reduce the probability of default.

SUGGESTIONS

1. Adjust Credit Limits Based on Default Risk:

To better balance risk and opportunity, it’s recommended to lower credit limits for high-risk customers while considering increasing limits for those with low-risk profiles. According to the model we have identified that higher credit limits are associated with a slightly reduced probability of default, making credit limit adjustments a valuable tool for managing risk.

2. Enhance Focus on Payment History:

Rigorous monitoring and early intervention strategies should be implemented for customers with severe payment delays (PAY_0). Since such delays are closely tied to higher default risks, offering reminders, support, or even temporary credit limit adjustments could mitigate the chances of default.

3. Include Age and Marital Status in Risk Assessment:

Incorporating age and marital status into risk assessments allows for more targeted interventions. Older customers have a slightly lower probability of default, and certain marital statuses may also pose risk. Customizing credit offers and interventions based on these factors could lead to better credit risk management.

4. Prioritise Recent Billing and Payment Patterns:

Customers with recent low payment amounts (PAY_AMT1) and high billing amounts (BILL_AMT1) should be closely monitored, as these patterns are strong predictors of defaulting. Prioritising such customers for additional checks and intervention can help reduce the likelihood of missed risks.

5. Improve Prediction for High-Risk Groups:

Efforts should be made to refine the model or introduce additional features to reduce false negatives, where actual defaulters are missed. Despite the model’s high accuracy (81.15%), the current number of false negatives (1470) suggests the need for enhanced prediction capabilities to better capture high-risk individuals.

CONCLUSION

The predictive model developed for Martins Bank achieves an accuracy of 81.15%, indicating reliable method for identifying non-defaulters. However, it struggles with predicting actual defaulters, as evidenced demonstrated a significant number of false negatives (1,470). Credit limit, age, gender, and payment history—particularly early payment delays (PAY_0)—are important variables that affect default risk. While the model performs well overall, there is room for improvement in detecting high-risk customers.

Martins Bank should modify credit limits according to risk profiles, emphasizing clients with unsatisfactory payment records, in order to improve credit risk management. Default risk can be further decreased by keeping a close eye on high-risk groups (those with high bills or poor recent payments).

Furthermore, in order to more accurately identify possible defaulters, the model must be improved to minimize false negatives. Martins Bank can ensure responsible lending and increased profitability by strengthening its data-driven credit system through the integration of these tactics.

APPENDIX

Step	R Code	Description
Load necessary libraries	library(dplyr)	Load the dplyr library for data manipulation.
Load the dataset	getwd() data <- read.csv("UCI_Credit_Card.csv")	Load the dataset from the working directory.
View the structure of the data	str(data)	Display the structure of the dataset.
Convert columns to factors	data$SEX <- as.factor(data$SEX) data$EDUCATION <- as.factor(data$EDUCATION) data$MARRIAGE <- as.factor(data$MARRIAGE) data$default.payment.next.month <- as.factor(data$default.payment.next.month)	Convert necessary columns to factors to prepare for modeling.
Split the data into training and testing sets	set.seed(123) sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.7, 0.3)) train <- data[sample,] test <- data[!sample,]	Randomly split the data into 70% training and 30% testing datasets.
Fit a Logistic Regression model	model <- glm(default.payment.next.month ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_0 + PAY_2 + PAY_3 + PAY_4 + PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT4 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6, data = train, family = binomial)	Fit a logistic regression model to predict default.payment.next.month using multiple predictors.
Model summary	summary(model)	Display a detailed summary of the logistic regression model.

Predict on test set	predictions <- predict(model, test, type = "response") predicted_classes <- ifelse(predictions > 0.5, 1, 0)	Predict the probability of default on the test set, classify based on threshold 0.5.
Confusion matrix	confusion_matrix <- table(test$default.payment.next.month, predicted_classes) print(confusion_matrix)	Create and display the confusion matrix to evaluate model performance.
Calculate accuracy	accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix) print(paste("Accuracy:", round(accuracy, 4)))	Calculate the overall accuracy of the model based on the confusion matrix.