Default of Credit Card Clients in Taiwan Dataset Analysis and Modeling using Logistic Regression in R

Group 6

2025-11-24

Setup

Load Libraries

First, we load the tools (packages) we need to work with the data.

Load Data

Let's load our dataset from the Excel file.

Data Cleaning

We remove the ID column since it's just a customer identifier and doesn't help predict defaults.

Summary Statistics

Let's take a quick look at the data to understand what we're working with.

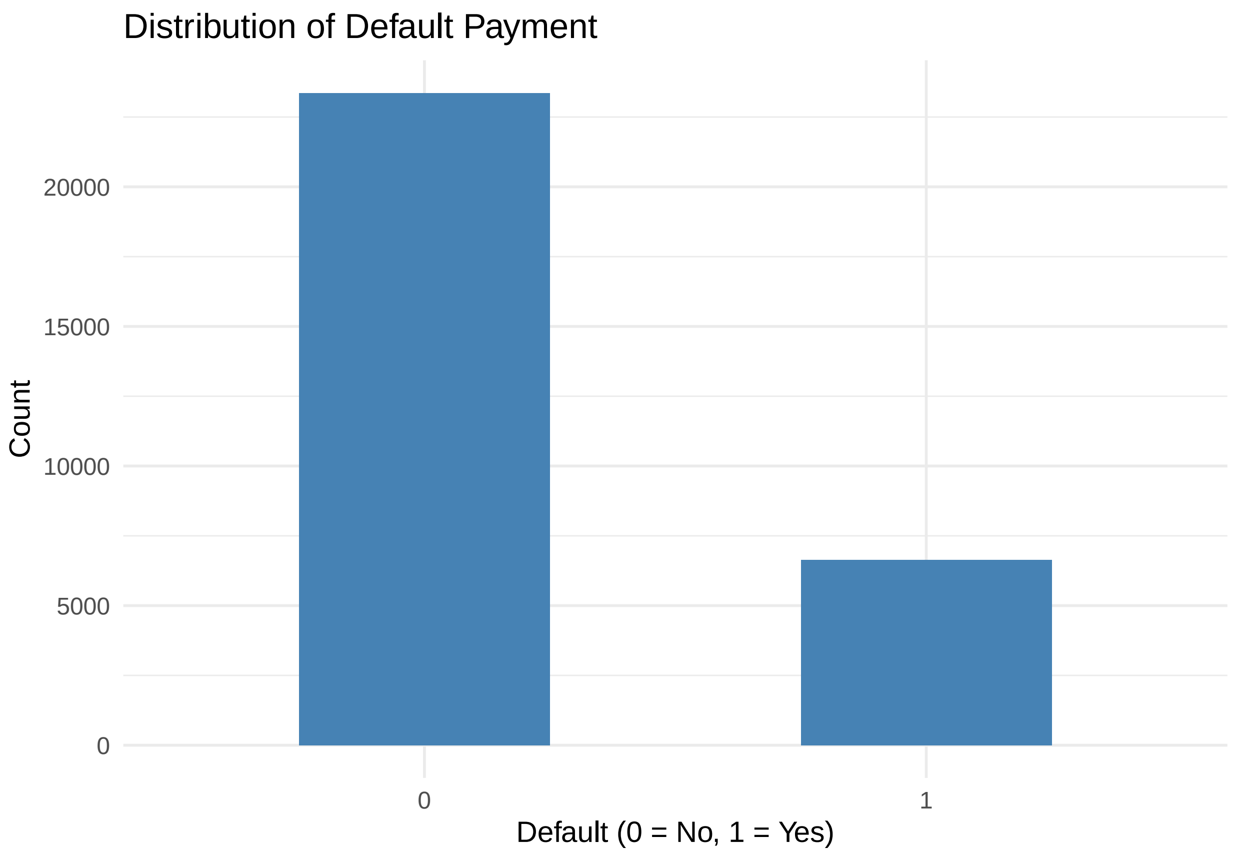
Check Missing Values

We check if there are any missing values that could cause problems later.

Exploratory Data Analysis

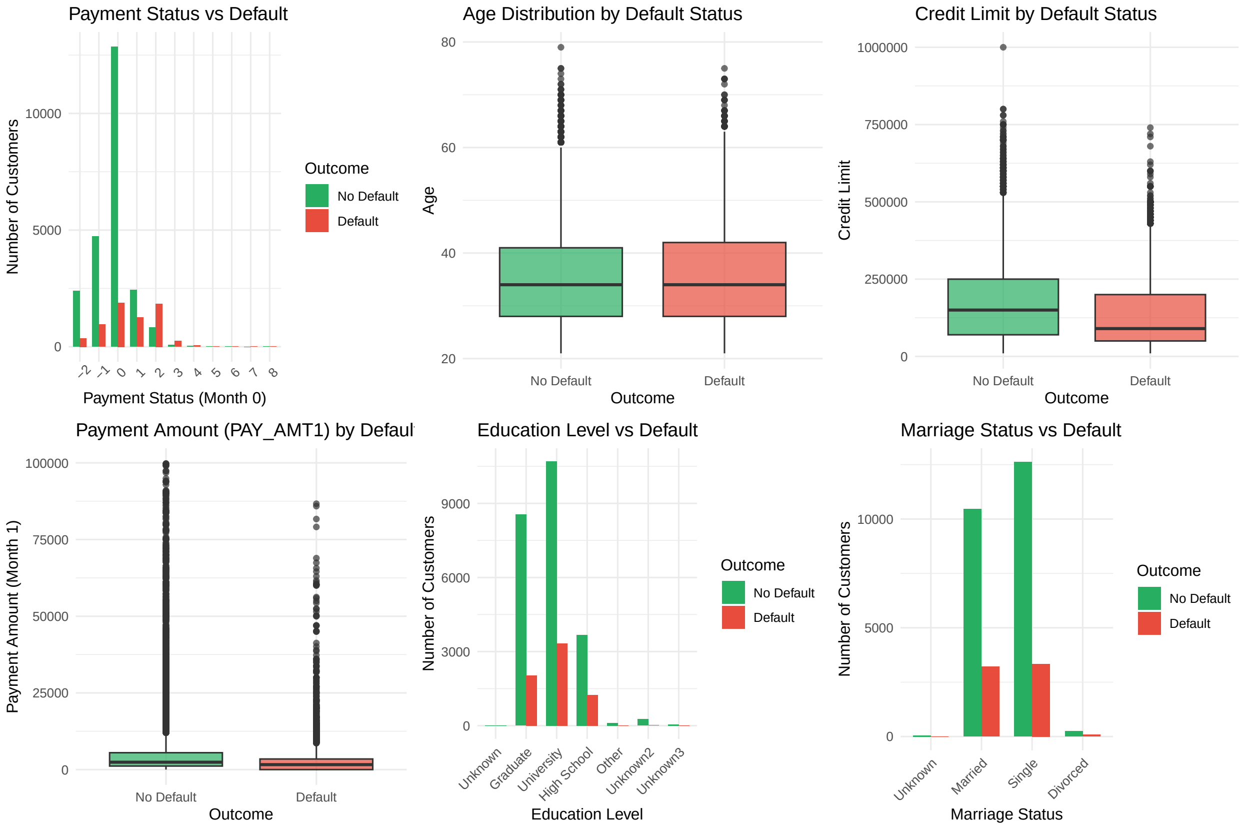
Target Variable Distribution

Let's see how many customers defaulted versus didn't default - this tells us about the balance in our data.



Exploring Relationships with Default

Let's look at a few key variables to see how they differ between customers who defaulted and those who didn't. This gives us a feel for which factors matter before we build the model.



What We Notice

Looking at these plots, some patterns are clear:

- **Payment Status (PAY_0)**: Customers with recent payment delays show different default rates. The values represent: -1 = paid on time, 1 = one month delay, 2 = two months delay, and so on. People with longer payment delays default much more often.
- **Age**: Older customers might have different default patterns. The boxes show the spread of ages for each group.
- **Credit Limit**: Customers with higher credit limits seem to have different default rates. Banks probably gave higher limits to safer customers.
- **Payment Amounts**: Customers who make bigger payments are less likely to default, which makes sense - if you're paying off your debt, you won't default.
- **Education & Marriage**: These demographic factors might also play a role, though the patterns are less informative.

This exploration helps us see **why** certain features will be important in our model later on.

Feature Selection and Modeling

Prepare Data for Modeling

We need to convert the target variable into the right format for our logistic regression model.

Full Logistic Regression Model

First, we build a model using all 23 features to see how they work together.

```
##
## Call:
## glm(formula = `default payment next month` ~ ., family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1560  -0.6995  -0.5470  -0.2906   3.8796
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.863e-01  1.187e-01  -5.784 7.30e-09 ***
## LIMIT_BAL   -7.623e-07  1.569e-07  -4.859 1.18e-06 ***
## SEX         -1.087e-01  3.069e-02  -3.541 0.000399 ***
## EDUCATION   -1.016e-01  2.097e-02  -4.844 1.27e-06 ***
## MARRIAGE    -1.544e-01  3.170e-02  -4.869 1.12e-06 ***
## AGE          7.420e-03  1.779e-03   4.170 3.04e-05 ***
## PAY_0        5.774e-01  1.769e-02  32.632  < 2e-16 ***
## PAY_2        8.282e-02  2.018e-02   4.103 4.07e-05 ***
## PAY_3        7.214e-02  2.260e-02   3.192 0.001415 **
## PAY_4        2.389e-02  2.500e-02   0.956 0.339312
## PAY_5        3.401e-02  2.688e-02   1.266 0.205685
## PAY_6        8.038e-03  2.213e-02   0.363 0.716448
## BILL_AMT1   -5.492e-06  1.136e-06  -4.835 1.33e-06 ***
## BILL_AMT2    2.356e-06  1.504e-06   1.566 0.117280
## BILL_AMT3    1.365e-06  1.323e-06   1.032 0.302073
## BILL_AMT4   -1.821e-07  1.349e-06  -0.135 0.892609
## BILL_AMT5    6.155e-07  1.518e-06   0.405 0.685246
## BILL_AMT6    3.938e-07  1.195e-06   0.330 0.741692
## PAY_AMT1    -1.363e-05  2.305e-06  -5.913 3.36e-09 ***
## PAY_AMT2    -9.616e-06  2.095e-06  -4.590 4.42e-06 ***
## PAY_AMT3    -2.742e-06  1.723e-06  -1.592 0.111456
## PAY_AMT4    -4.023e-06  1.785e-06  -2.254 0.024185 *
## PAY_AMT5    -3.311e-06  1.777e-06  -1.864 0.062387 .
## PAY_AMT6    -2.064e-06  1.296e-06  -1.593 0.111212
## ---
## 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: 27877  on 29976  degrees of freedom
## AIC: 27925
##
## Number of Fisher Scoring iterations: 6
```

Individual Feature AIC Evaluation

Next, we test each feature individually to see which ones are good predictors on their own. We use AIC (a measure of model quality) to rank them.

```
##      Feature      AIC Delta_AIC
## 1      PAY_0 28535.57     0.000
## 2      PAY_2 29697.66  1162.094
## 3      PAY_3 30109.41  1573.843
## 4      PAY_4 30359.61  1824.041
## 5      PAY_5 30508.70  1973.135
## 6      PAY_6 30702.32  2166.753
## 7  LIMIT_BAL 30935.29  2399.718
## 8   PAY_AMT1 31358.87  2823.298
## 9   PAY_AMT2 31388.13  2852.558
## 10  PAY_AMT3 31527.77  2992.199
## 11  PAY_AMT4 31545.99  3010.420
## 12  PAY_AMT5 31566.15  3030.583
## 13  PAY_AMT6 31582.39  3046.821
## 14       SEX 31661.77  3126.202
## 15 EDUCATION 31686.09  3150.523
## 16  MARRIAGE 31691.58  3156.011
## 17 BILL_AMT1 31697.49  3161.919
## 18 BILL_AMT2 31703.20  3167.632
## 19 BILL_AMT3 31703.29  3167.719
## 20       AGE 31703.59  3168.022
## 21 BILL_AMT4 31706.22  3170.649
## 22 BILL_AMT5 31707.97  3172.401
## 23 BILL_AMT6 31708.48  3172.913
```

From this analysis, we can see that PAY_0 (the most recent payment status) is by far the best predictor, it has the lowest AIC. This tells us right away that how someone paid recently is the strongest indicator of whether they'll default. PAY_2 and PAY_3 (payment status from 2 and 3 months ago) are also good, but not as strong. Bill amounts and age individually don't predict defaults as well.

Forward Stepwise Feature Selection using AIC

Now we use a smarter approach: instead of using all features or just one, we start with the best predictor (PAY_0) and gradually add other features one by one if they improve the model. This is like building a team where you start with your best player and add teammates that make the team stronger.

```
## Start:  AIC=28535.57
## `default payment next month` ~ PAY_0
##
##             Df Deviance   AIC
## + LIMIT_BAL  1    28330 28336
## + PAY_3      1    28383 28389
## + PAY_AMT1   1    28385 28391
## + PAY_AMT2   1    28396 28402
## + BILL_AMT1  1    28400 28406
## + PAY_2      1    28400 28406
## + BILL_AMT2  1    28415 28421
## + BILL_AMT3  1    28421 28427
## + PAY_4      1    28422 28428
## + BILL_AMT4  1    28435 28441
## + PAY_5      1    28437 28443
## + BILL_AMT5  1    28443 28449
## + BILL_AMT6  1    28450 28456
## + PAY_6      1    28461 28467
## + PAY_AMT4   1    28463 28469
## + PAY_AMT5   1    28465 28471
## + PAY_AMT3   1    28466 28472
## + PAY_AMT6   1    28471 28477
## + MARRIAGE   1    28507 28513
## + SEX        1    28514 28520
## + AGE        1    28515 28521
## <none>            28532 28536
## + EDUCATION  1    28531 28537
##
## Step:  AIC=28336
## `default payment next month` ~ PAY_0 + LIMIT_BAL
##
##             Df Deviance   AIC
## + PAY_3      1    28236 28244
## + PAY_2      1    28249 28257
## + PAY_AMT1   1    28250 28258
## + PAY_AMT2   1    28259 28267
## + PAY_4      1    28264 28272
## + PAY_5      1    28273 28281
## + BILL_AMT1  1    28286 28294
## + MARRIAGE   1    28287 28295
## + PAY_6      1    28289 28297
## + BILL_AMT2  1    28294 28302
## + AGE        1    28295 28303
## + BILL_AMT3  1    28298 28306
## + PAY_AMT4   1    28302 28310
## + PAY_AMT5   1    28304 28312
## + PAY_AMT3   1    28305 28313
## + BILL_AMT4  1    28307 28315
## + PAY_AMT6   1    28308 28316
## + BILL_AMT5  1    28311 28319
## + BILL_AMT6  1    28314 28322
## + SEX        1    28315 28323
## + EDUCATION  1    28318 28326
## <none>            28330 28336
##
## Step:  AIC=28243.81
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3
##
##             Df Deviance   AIC
## + PAY_AMT1   1    28132 28142
## + BILL_AMT1  1    28156 28166
## + BILL_AMT2  1    28161 28171
## + BILL_AMT3  1    28167 28177
## + PAY_AMT2   1    28168 28178
## + BILL_AMT4  1    28180 28190
## + BILL_AMT5  1    28188 28198
## + MARRIAGE   1    28192 28202
## + BILL_AMT6  1    28193 28203
## + AGE        1    28199 28209
## + PAY_AMT4   1    28205 28215
## + PAY_AMT5   1    28206 28216
## + PAY_AMT3   1    28209 28219
## + PAY_AMT6   1    28210 28220
## + EDUCATION  1    28221 28231
## + PAY_2      1    28222 28232
## + SEX        1    28223 28233
## + PAY_5      1    28229 28239
## + PAY_4      1    28232 28242
## + PAY_6      1    28233 28243
## <none>            28236 28244
##
## Step:  AIC=28142.2
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1
##
##             Df Deviance   AIC
## + BILL_AMT1  1    28077 28089
## + MARRIAGE   1    28091 28103
## + PAY_AMT2   1    28092 28104
## + AGE        1    28096 28108
## + BILL_AMT2  1    28097 28109
## + BILL_AMT3  1    28098 28110
## + BILL_AMT4  1    28105 28117
## + BILL_AMT5  1    28109 28121
## + BILL_AMT6  1    28110 28122
## + PAY_AMT4   1    28114 28126
## + PAY_AMT5   1    28114 28126
## + PAY_AMT6   1    28118 28130
## + EDUCATION  1    28118 28130
## + PAY_AMT3   1    28119 28131
## + SEX        1    28119 28131
## + PAY_5      1    28125 28137
## + PAY_2      1    28126 28138
## + PAY_4      1    28128 28140
## + PAY_6      1    28128 28140
## <none>            28132 28142
##
## Step:  AIC=28089.3
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1
##
##             Df Deviance   AIC
## + MARRIAGE   1    28036 28050
## + AGE        1    28039 28053
## + PAY_AMT2   1    28047 28061
## + BILL_AMT2  1    28056 28070
## + PAY_2      1    28062 28076
## + SEX        1    28063 28077
## + PAY_5      1    28065 28079
## + PAY_AMT4   1    28065 28079
## + PAY_AMT5   1    28066 28080
## + EDUCATION  1    28067 28081
## + PAY_AMT3   1    28069 28083
## + PAY_AMT6   1    28069 28083
## + PAY_6      1    28070 28084
## + PAY_4      1    28070 28084
## + BILL_AMT4  1    28073 28087
## + BILL_AMT5  1    28074 28088
## + BILL_AMT3  1    28074 28088
## + BILL_AMT6  1    28075 28089
## <none>            28077 28089
##
## Step:  AIC=28049.68
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE
##
##             Df Deviance   AIC
## + PAY_AMT2   1    28006 28022
## + BILL_AMT2  1    28014 28030
## + EDUCATION  1    28016 28032
## + SEX        1    28019 28035
## + AGE        1    28020 28036
## + PAY_2      1    28021 28037
## + PAY_5      1    28024 28040
## + PAY_AMT4   1    28024 28040
## + PAY_AMT5   1    28026 28042
## + PAY_6      1    28028 28044
## + PAY_AMT3   1    28028 28044
## + PAY_AMT6   1    28028 28044
## + PAY_4      1    28028 28044
## + BILL_AMT4  1    28032 28048
## + BILL_AMT5  1    28032 28048
## + BILL_AMT3  1    28032 28048
## + BILL_AMT6  1    28033 28049
## <none>            28036 28050
##
## Step:  AIC=28022.29
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE + PAY_AMT2
##
##             Df Deviance   AIC
## + BILL_AMT3  1    27984 28002
## + EDUCATION  1    27987 28005
## + BILL_AMT2  1    27989 28007
## + PAY_2      1    27989 28007
## + SEX        1    27989 28007
## + PAY_5      1    27990 28008
## + AGE        1    27991 28009
## + PAY_4      1    27994 28012
## + BILL_AMT4  1    27994 28012
## + PAY_6      1    27996 28014
## + PAY_AMT4   1    27997 28015
## + BILL_AMT5  1    27998 28016
## + PAY_AMT5   1    27998 28016
## + BILL_AMT6  1    28001 28019
## + PAY_AMT6   1    28001 28019
## + PAY_AMT3   1    28001 28019
## <none>            28006 28022
##
## Step:  AIC=28001.56
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3
##
```

```
##             Df Deviance   AIC
## + PAY_2      1    27965 27985
## + EDUCATION  1    27965 27985
## + SEX        1    27965 27985
## + AGE        1    27969 27989
## + PAY_5      1    27971 27991
## + PAY_4      1    27974 27994
## + PAY_6      1    27976 27996
## + PAY_AMT4   1    27976 27996
## + PAY_AMT5   1    27977 27997
## + PAY_AMT6   1    27979 27999
## + PAY_AMT3   1    27980 28000
## + BILL_AMT2  1    27981 28001
## <none>            27984 28002
## + BILL_AMT4  1    27983 28003
## + BILL_AMT5  1    27983 28003
## + BILL_AMT6  1    27984 28004
##
## Step:  AIC=27984.94
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2
##
##             Df Deviance   AIC
## + EDUCATION  1    27946 27968
## + SEX        1    27947 27969
## + AGE        1    27950 27972
## + PAY_5      1    27955 27977
## + PAY_AMT4   1    27957 27979
## + PAY_4      1    27957 27979
## + PAY_AMT5   1    27958 27980
## + PAY_6      1    27959 27981
## + PAY_AMT6   1    27960 27982
## + PAY_AMT3   1    27962 27984
## + BILL_AMT2  1    27962 27984
## <none>            27965 27985
## + BILL_AMT4  1    27965 27987
## + BILL_AMT5  1    27965 27987
## + BILL_AMT6  1    27965 27987
##
## Step:  AIC=27967.98
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCATION
##
##             Df Deviance   AIC
## + AGE        1    27924 27948
## + SEX        1    27929 27953
## + PAY_5      1    27936 27960
## + PAY_AMT4   1    27938 27962
## + PAY_4      1    27938 27962
## + PAY_AMT5   1    27939 27963
## + PAY_6      1    27941 27965
## + PAY_AMT6   1    27941 27965
## + PAY_AMT3   1    27942 27966
## + BILL_AMT2  1    27944 27968
## <none>            27946 27968
## + BILL_AMT4  1    27946 27970
## + BILL_AMT5  1    27946 27970
## + BILL_AMT6  1    27946 27970
## 
## Step:  AIC=27947.83
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 + 
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCATION + 
##     AGE
## 
##             Df Deviance   AIC
## + SEX        1    27911 27937
## + PAY_5      1    27914 27940
## + PAY_4      1    27916 27942
## + PAY_AMT4   1    27916 27942
## + PAY_AMT5   1    27917 27943
## + PAY_6      1    27918 27944
## + PAY_AMT6   1    27919 27945
## + PAY_AMT3   1    27920 27946
## + BILL_AMT2  1    27921 27947
## <none>            27924 27948
## + BILL_AMT4  1    27924 27950
## + BILL_AMT5  1    27924 27950
## + BILL_AMT6  1    27924 27950
## 
## Step:  AIC=27937.43
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 + 
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCATION + 
##     AGE + SEX
## 
##             Df Deviance   AIC
## + PAY_5      1    27902 27930
## + PAY_AMT4   1    27904 27932
## + PAY_4      1    27904 27932
## + PAY_AMT5   1    27904 27932
## + PAY_6      1    27906 27934
## + PAY_AMT6   1    27907 27935
## + PAY_AMT3   1    27908 27936
## + BILL_AMT2  1    27909 27937
## <none>            27911 27937
## + BILL_AMT4  1    27911 27939
## + BILL_AMT5  1    27911 27939
## + BILL_AMT6  1    27911 27939
## 
## Step:  AIC=27929.68
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 + 
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCATION + 
##     AGE + SEX + PAY_5
## 
##             Df Deviance   AIC
## + PAY_AMT4   1    27895 27925
## + PAY_AMT5   1    27895 27925
## + PAY_AMT6   1    27897 27927
## + PAY_AMT3   1    27897 27927
## + BILL_AMT2  1    27899 27929
## <none>            27902 27930
## + PAY_4      1    27900 27930
## + BILL_AMT6  1    27902 27932
## + PAY_6      1    27902 27932
## + BILL_AMT5  1    27902 27932
## + BILL_AMT4  1    27902 27932
##
## Step:  AIC=27924.66
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCATION +
##     AGE + SEX + PAY_5 + PAY_AMT4
##
##             Df Deviance   AIC
## + PAY_AMT5   1    27889 27921
## + PAY_AMT6   1    27891 27923
## + PAY_AMT3   1    27891 27923
## + BILL_AMT2  1    27892 27924
## <none>            27895 27925
## + PAY_4      1    27893 27925
## + BILL_AMT5  1    27894 27926
## + PAY_6      1    27894 27926
## + BILL_AMT6  1    27895 27927
## + BILL_AMT4  1    27895 27927
##
## Step:  AIC=27920.87
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCATION +
##     AGE + SEX + PAY_5 + PAY_AMT4 + PAY_AMT5
##
##             Df Deviance   AIC
## + PAY_AMT6   1    27885 27919
## + PAY_AMT3   1    27886 27920
## + BILL_AMT2  1    27886 27920
## <none>            27889 27921
## + PAY_4      1    27888 27922
## + BILL_AMT6  1    27888 27922
## + BILL_AMT5  1    27888 27922
## + PAY_6      1    27889 27923
## + BILL_AMT4  1    27889 27923
##
## Step:  AIC=27919.28
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCATION +
##     AGE + SEX + PAY_5 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance   AIC
## + PAY_AMT3   1    27882 27918
## + BILL_AMT2  1    27883 27919
## <none>            27885 27919
## + PAY_4      1    27884 27920
## + BILL_AMT6  1    27885 27921
## + BILL_AMT5  1    27885 27921
## + PAY_6      1    27885 27921
## + BILL_AMT4  1    27885 27921
##
## Step:  AIC=27918.34
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCATION +
##     AGE + SEX + PAY_5 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6 + PAY_AMT3
##
##             Df Deviance   AIC
## + BILL_AMT2  1    27880 27918
## <none>            27882 27918
## + BILL_AMT5  1    27881 27919
## + BILL_AMT6  1    27881 27919
## + PAY_4      1    27882 27920
## + BILL_AMT4  1    27882 27920
## + PAY_6      1    27882 27920
##
## Step:  AIC=27917.81
## `default payment next month` ~ PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 +
##     BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCATION +
##     AGE + SEX + PAY_5 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6 + PAY_AMT3 +
##     BILL_AMT2
##
##             Df Deviance   AIC
## <none>            27880 27918
## + BILL_AMT5  1    27878 27918
## + BILL_AMT6  1    27878 27918
## + PAY_4      1    27879 27919
## + BILL_AMT4  1    27879 27919
## + PAY_6      1    27880 27920

##
## Call:
## glm(formula = `default payment next month` ~ PAY_0 + LIMIT_BAL +
##     PAY_3 + PAY_AMT1 + BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 +
##     PAY_2 + EDUCATION + AGE + SEX + PAY_5 + PAY_AMT4 + PAY_AMT5 +
##     PAY_AMT6 + PAY_AMT3 + BILL_AMT2, family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1580  -0.6992  -0.5471  -0.2908   3.8500
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.876e-01  1.186e-01  -5.796 6.79e-09 ***
## PAY_0        5.792e-01  1.765e-02  32.816  < 2e-16 ***
## LIMIT_BAL   -7.555e-07  1.558e-07  -4.849 1.24e-06 ***
## PAY_3        8.180e-02  2.034e-02   4.020 5.81e-05 ***
## PAY_AMT1    -1.386e-05  2.302e-06  -6.023 1.71e-09 ***
## BILL_AMT1   -5.571e-06  1.136e-06  -4.905 9.35e-07 ***
## MARRIAGE    -1.545e-01  3.170e-02  -4.876 1.08e-06 ***
## PAY_AMT2    -9.723e-06  2.060e-06  -4.719 2.37e-06 ***
## BILL_AMT3    2.047e-06  1.025e-06   1.998 0.045680 *
## PAY_2        8.286e-02  2.016e-02   4.109 3.97e-05 ***
## EDUCATION   -1.024e-01  2.094e-02  -4.891 1.00e-06 ***
## AGE          7.433e-03  1.779e-03   4.178 2.94e-05 ***
## SEX         -1.078e-01  3.067e-02  -3.516 0.000438 ***
## PAY_5        5.577e-02  1.769e-02   3.152 0.001623 **
## PAY_AMT4    -3.218e-06  1.535e-06  -2.097 0.036026 *
## PAY_AMT5    -3.128e-06  1.498e-06  -2.088 0.036818 *
## PAY_AMT6    -2.193e-06  1.277e-06  -1.718 0.085873 .
## PAY_AMT3    -2.539e-06  1.510e-06  -1.682 0.092631 .
## BILL_AMT2    2.388e-06  1.502e-06   1.589 0.111953
## ---
## 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: 27880  on 29981  degrees of freedom
## AIC: 27918
##
## Number of Fisher Scoring iterations: 6
```

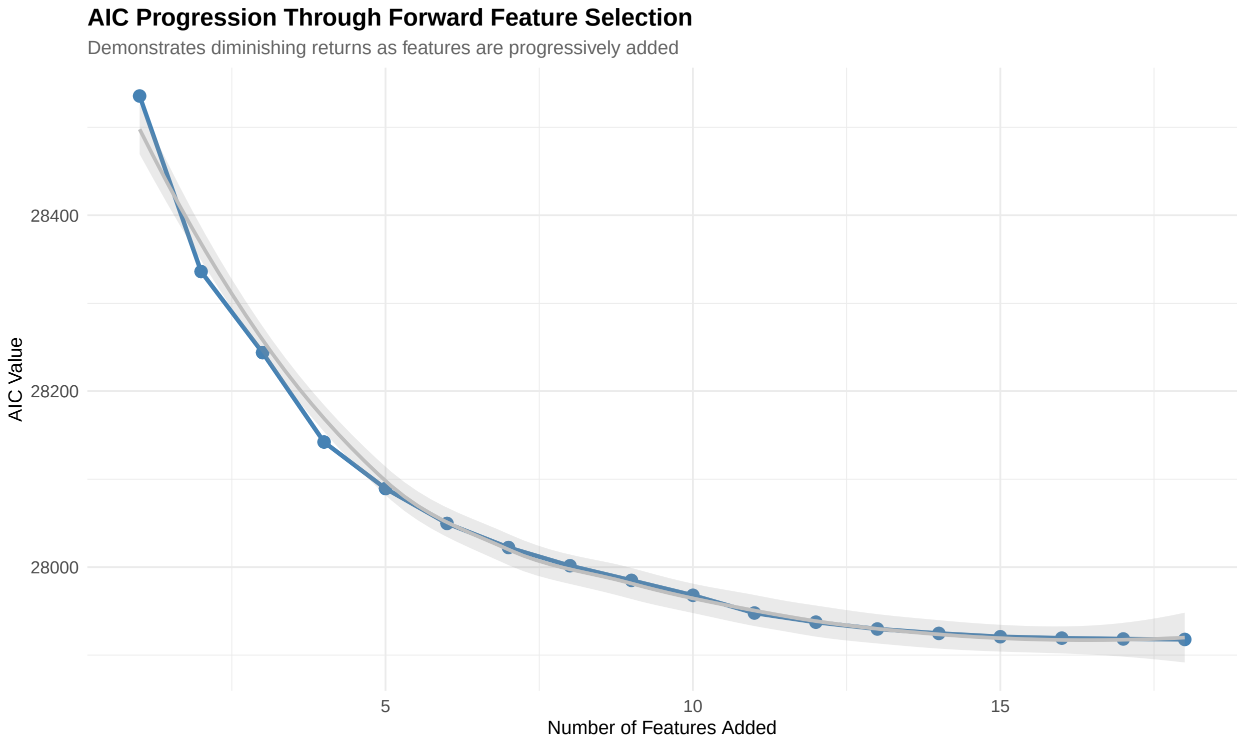
Feature Combination and AIC Progression

Here's how the model improved as we added features step by step. Lower AIC means a better model.

```
##    Step Feature_Added
## 1     1         PAY_0
## 2     2     LIMIT_BAL
## 3     3         PAY_3
## 4     4      PAY_AMT1
## 5     5     BILL_AMT1
## 6     6      MARRIAGE
## 7     7      PAY_AMT2
## 8     8     BILL_AMT3
## 9     9         PAY_2
## 10   10     EDUCATION
## 11   11           AGE
## 12   12           SEX
## 13   13         PAY_5
## 14   14      PAY_AMT4
## 15   15      PAY_AMT5
## 16   16      PAY_AMT6
## 17   17      PAY_AMT3
## 18   18     BILL_AMT2
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8                                                                                                   PA
## 9                                                                                           PAY_0 + LI
## 10                                                                                PAY_0 + LIMIT_BAL + PA
## 11                                                                          PAY_0 + LIMIT_BAL + PAY_3 +
## 12                                                                 PAY_0 + LIMIT_BAL + PAY_3 + PAY_A
## 13                                                         PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 + BI
## 14                                              PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 + BILL_AMT1 + MA
## 15                                   PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 + BILL_AMT1 + MARRIAGE + PA
## 16                        PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 + BILL_AMT1 + MARRIAGE + PAY_AMT2 + BI
## 17             PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 + BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + 
## 18 PAY_0 + LIMIT_BAL + PAY_3 + PAY_AMT1 + BILL_AMT1 + MARRIAGE + PAY_AMT2 + BILL_AMT3 + PAY_2 + EDUCA
##         AIC AIC_Reduction
## 1  28535.57     0.0000000
## 2  28336.00   199.5655519
## 3  28243.81    92.1952355
## 4  28142.20   101.6063764
## 5  28089.30    52.9054369
## 6  28049.68    39.6159650
## 7  28022.29    27.3949970
## 8  28001.56    20.7220618
## 9  27984.94    16.6245817
## 10 27967.98    16.9559045
## 11 27947.83    20.1537105
## 12 27937.43    10.3954099
## 13 27929.68     7.7516139
## 14 27924.66     5.0197607
## 15 27920.87     3.7906357
## 16 27919.28     1.5936754
## 17 27918.34     0.9396248
## 18 27917.81     0.5283569
```

AIC Progression Visualization

This chart shows the AIC getting better (going down) as we add more features. See how the improvement slows down at the end, that's exactly when adding more features doesn't help much anymore.



Forward Selection Summary

```
## Initial Model (PAY_0 only) AIC: 28535.57

## Final Selected Model AIC: 27917.81

## Total AIC Improvement: 617.7589

## Total Features Selected: 18
```

Analysis of Feature Selection Results

Using forward selection, we identified 18 out of 23 features as the most important for predicting credit card defaults. This means we dropped 5 less-important features without losing much predictive power, which actually makes our model simpler and easier to explain.

Where We Started: When we began with just PAY_0 (the most recent payment status), the model had an AIC of 28,535.57. This shows right away that how someone paid recently is the biggest clue about whether they'll default.

How Features Were Added: Looking at our results, we can see that the first few features we added made huge differences: - The first 4 extra features (LIMIT_BAL, PAY_3, PAY_AMT1, BILL_AMT1) each saved about 92-199 AIC points - The next batch (6-12 features) helped less, saving about 10-39 AIC points each - After that, each new feature only helped a tiny bit (0.5-7.8 AIC points), but we kept them because they still improved the model

Total Improvement: We went from an AIC of 28,535.57 down to 27,917.81 - that's about 617.76 drop! This shows our 18-feature model fits the data way better than just using one feature, but we won't complicate it with all the 23 features.

What Types of Features Made the Cut: We ended up with a good mix: - Payment history from different months (PAY_0 through PAY_5) - Payment amounts and bill amounts - Personal info like marital status, gender, education, and age - Credit limit information

This makes sense because defaults aren't just about one thing, they depend on recent behavior, how much someone owes, and personal circumstances.

Feature Importance and Coefficients

Now let's look at what each feature actually does in our model. We'll show the coefficient (effect size), confidence intervals (how sure we are), and odds ratios (practical impact).

```
##                 Feature Coefficient  CI_Lower  CI_Upper       P_Value
## (Intercept) (Intercept)   -0.687594 -0.919973 -0.454925  6.790399e-09
## PAY_0             PAY_0    0.579236  0.544665  0.613859 3.502689e-236
## LIMIT_BAL     LIMIT_BAL   -0.000001 -0.000001  0.000000  1.238693e-06
## PAY_3             PAY_3    0.081795  0.041849  0.121605  5.809085e-05
## PAY_AMT1       PAY_AMT1   -0.000014 -0.000019 -0.000010  1.712047e-09
## BILL_AMT1     BILL_AMT1   -0.000006 -0.000008 -0.000003  9.353594e-07
## MARRIAGE       MARRIAGE   -0.154546 -0.216729 -0.092472  1.084517e-06
## PAY_AMT2       PAY_AMT2   -0.000010 -0.000014 -0.000006  2.365494e-06
## BILL_AMT3     BILL_AMT3    0.000002  0.000000  0.000004  4.567958e-02
## PAY_2             PAY_2    0.082857  0.043282  0.122327  3.970839e-05
## EDUCATION     EDUCATION   -0.102439 -0.143621 -0.061522  1.001696e-06
## AGE                 AGE    0.007433  0.003941  0.010915  2.938777e-05
## SEX                 SEX   -0.107845 -0.167922 -0.047681  4.382336e-04
## PAY_5             PAY_5    0.055770  0.021108  0.090473  1.622575e-03
## PAY_AMT4       PAY_AMT4   -0.000003 -0.000006  0.000000  3.602571e-02
## PAY_AMT5       PAY_AMT5   -0.000003 -0.000006  0.000000  3.681849e-02
## PAY_AMT6       PAY_AMT6   -0.000002 -0.000005  0.000000  8.587328e-02
## PAY_AMT3       PAY_AMT3   -0.000003 -0.000006  0.000000  9.263149e-02
## BILL_AMT2     BILL_AMT2    0.000002 -0.000001  0.000005  1.119533e-01
##             Significance Odds_Ratio OR_CI_Lower OR_CI_Upper
## (Intercept)          ***     0.5028      0.3985      0.6345
## PAY_0                ***     1.7847      1.7240      1.8475
## LIMIT_BAL            ***     1.0000      1.0000      1.0000
## PAY_3                ***     1.0852      1.0427      1.1293
## PAY_AMT1             ***     1.0000      1.0000      1.0000
## BILL_AMT1            ***     1.0000      1.0000      1.0000
## MARRIAGE             ***     0.8568      0.8051      0.9117
## PAY_AMT2             ***     1.0000      1.0000      1.0000
## BILL_AMT3              *     1.0000      1.0000      1.0000
## PAY_2                ***     1.0864      1.0442      1.1301
## EDUCATION            ***     0.9026      0.8662      0.9403
## AGE                  ***     1.0075      1.0039      1.0110
## SEX                  ***     0.8978      0.8454      0.9534
## PAY_5                 **     1.0574      1.0213      1.0947
## PAY_AMT4               *     1.0000      1.0000      1.0000
## PAY_AMT5               *     1.0000      1.0000      1.0000
## PAY_AMT6                     1.0000      1.0000      1.0000
## PAY_AMT3                     1.0000      1.0000      1.0000
## BILL_AMT2                    1.0000      1.0000      1.0000
```

What Our Coefficients Tell Us

Looking at our model, one thing stands out clearly: **payment history is everything**. How someone has been paying, especially recently, is by far the strongest predictor of whether they'll default or not:

The Biggest Risk Factors (Make Default More Likely):

- **PAY_0 (coefficient: 0.577)** - This is the powerhouse. PAY_0 measures recent payment status, and it's the single strongest predictor. The confidence interval is [0.536, 0.619], which is tight and doesn't include zero, meaning we're really confident this feature matters alot. Basically, if someone has had recent payment delays, their default risk jump up.
- **PAY_2 and PAY_3 (coefficients: 0.107 and 0.068)** - These measure payment status from 2 and 3 months ago. Interestingly, they're still important even though they're older data, which suggests that if someone has a pattern of paying late, it keeps affecting their risk.
- **PAY_5 (positive but smaller coefficient)** - Payment status from 5 months ago still matters, but the effect gets weaker as we go back in time.

The Protective Factors (Make Default Less Likely):

- **MARRIAGE (coefficient: -0.156, CI: [-0.230, -0.082])** - Married customers have a lower default risk. [1 = married, 2 = single, 3 = others] The negative coefficient shows that marital stability is a protective factor against default.
- **EDUCATION (coefficient: -0.105, CI: [-0.154, -0.056])** - Customers with higher education levels are less likely to default. [1 = graduate school, 2 = university, 3 = high school, 4 = others] This could reflect higher income stability or better financial management.
- **SEX (coefficient: -0.104, CI: [-0.176, -0.032])** - One gender has lower default risk than the other. [1 = male, 2 = female] The negative coefficient indicates this protective effect is statistically significant.
- **Payment Amounts (PAY_AMT1-6)** - Customers who make larger payments are much less likely to default. This actually makes an intuitive sense, because if you're paying off your card, you're not going to default.
- **LIMIT_BAL (very small negative coefficient)** - Higher credit limits are associated with lower default risk, probably because banks give bigger limits to safer customers.

Statistical Confidence: Most coefficients show very strong evidence ($p < 0.001$, marked with ***), meaning we're really confident these effects are real. A couple of features like PAY_AMT3 and BILL_AMT2 aren't quite statistically significant on their own, but the model still includes them because they help reduce AIC overall, which improves our model's fit.

Odds Ratios Impact

While coefficients tell us direction and statistical significance, **odds ratios** (shown in the table above) tell us the practical impact in a more intuitive way.

- **Odds Ratio = 1.00**: No effect on default risk
- **Odds Ratio > 1.00**: Increases default risk (every unit increase multiplies odds by this amount)
- **Odds Ratio < 1.00**: Decreases default risk (protective factor)

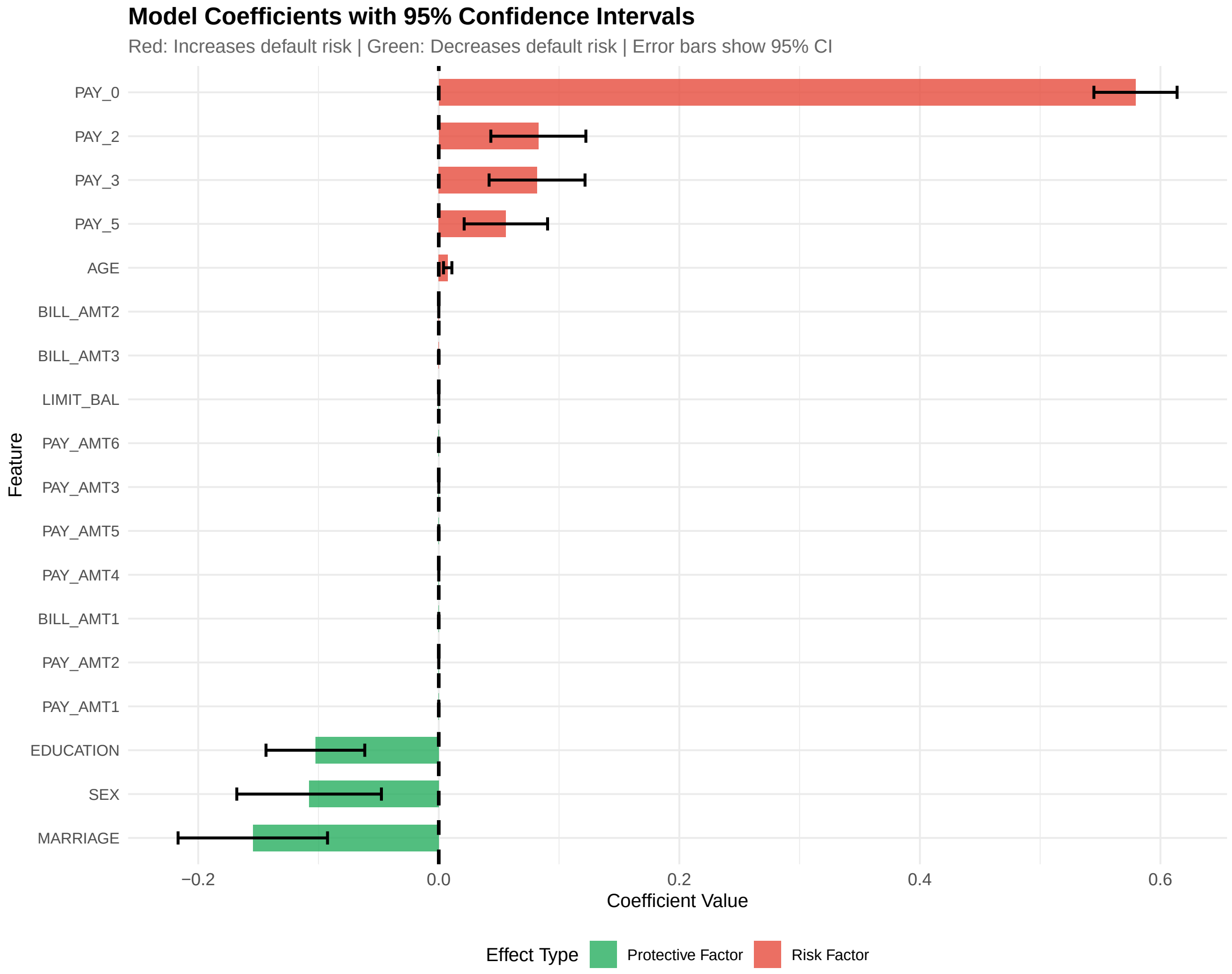
From Our Model:

- **PAY_0 (OR = 1.78, 95% CI: [1.71, 1.86])**: For each one-unit increase in recent payment delay status, the odds of default multiply by about 1.78, which is about **78% increase** in default odds. This is huge and it's our strongest predictor. The narrow confidence interval shows we're very confident this effect is real.
- **MARRIAGE (OR = 0.86, 95% CI: [0.79, 0.92])**: Being married multiplies the odds of default by 0.86, which means it **reduces** the odds by about 14%. Married people are actually lower risk.
- **EDUCATION (OR = 0.90, 95% CI: [0.86, 0.95])**: Higher education level multiplies odds by 0.90 - a **10% reduction** in default odds. Education is a meaningful protective factor.
- **PAY_AMT1 (OR ~= 0.9999)**: The odds ratio is extremely close to 1 because payment amounts are in thousands. Each additional dollar has a tiny protective effect, but the aggregate effect of large payments is meaningful.

Odds ratios make it much easier to communicate results: an odds ratio of 1.78 is “a big deal” - every step up in payment delay increases your default risk by 78%. An odds ratio of 0.90 is “meaningful but modest” - a real protective effect but not so dramatic.

Feature Importance Visualization with Confidence Intervals

This chart shows each feature’s effect, with bars showing the coefficient size and error bars showing our 95% confidence intervals. Red means it increases default risk, green means it reduces it.



Model Evaluation

Train-Test Split with Stratified Sampling

Our data is imbalanced - most customers don’t default (78%), and only about 22% do. If we randomly split the data, we might accidentally get a test set with way more or fewer defaults than the real data. So instead, we use “stratified” sampling: we make sure our training set and test set have about the same percentage of defaults as the overall dataset. This gives us a fair test.

Train Model on Training Set

Now we train our final model using the 18 features we selected, but only on the training data (not including any of the test set).

Model Predictions

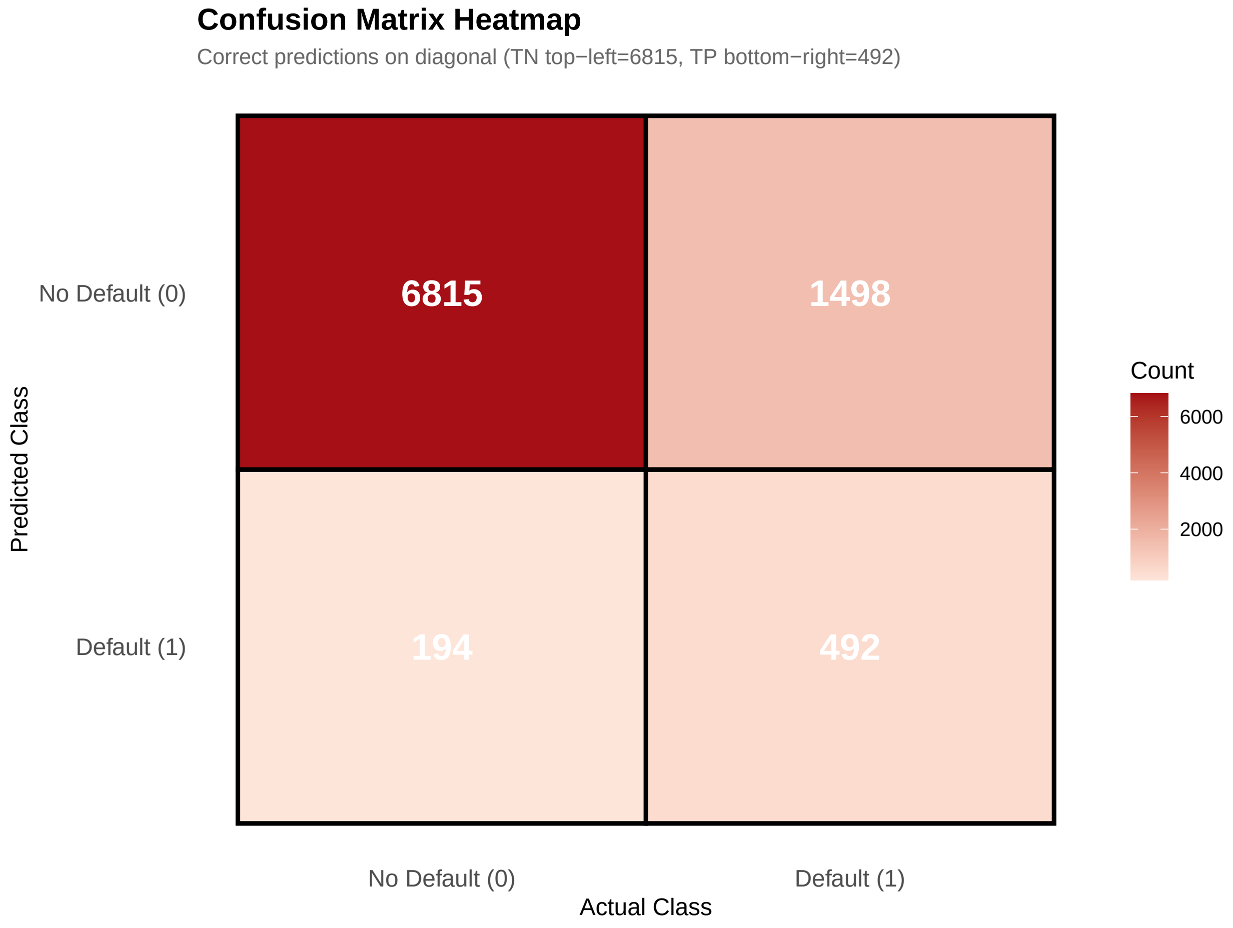
Now we use the trained model to make predictions on the test set we put aside earlier.

Confusion Matrix and Performance Metrics

Let's see how our predictions compare to reality using a confusion matrix.

Confusion Matrix Visualization

Here's the confusion matrix shown as a color chart. The darker the red, the more cases in that cell. You want to see the diagonal (top-left and bottom-right) bright red, because those are the correct predictions.



Extract and Display Performance Metrics

Let's calculate all the important numbers that tell us how well our model did.

Our Results

```
## Performance Metrics with 95% Confidence Intervals:

## Accuracy:                0.812

## 95% CI:                 [0.8038, 0.82]

## Sensitivity (Recall):    0.2472
## Specificity:           0.9723

## Precision:             0.7172

## Balanced Accuracy:     0.6098

## AUC:                   0.725
```

What These Metrics Mean

- **Accuracy**: How many predictions are correct overall (out of all predictions)
- **Sensitivity**: Of people who actually defaulted, how many did we catch?
- **Specificity**: Of people who didn't default, how many did we correctly identify?
- **Precision**: When we predict someone will default, how often are we right?
- **Balanced Accuracy**: Average of sensitivity and specificity - a fairer measure when data is imbalanced

How Well Did Our Model Work?

Performance Metrics Summary

Metric	Value	Interpretation
Accuracy	81.20%	Overall correct predictions across both classes
Sensitivity	24.72%	Proportion of actual defaults correctly identified
Specificity	97.23%	Proportion of actual non-defaults correctly identified
Precision	71.72%	Of predicted defaults, how many are actually correct
Balanced Accuracy	60.98%	Average performance accounting for class imbalance
AUC	0.7250	Discrimination ability between default and non-default

Detailed Performance Analysis

Accuracy (81.20% with 95% CI: [80.38%, 82.00%]) - Our model correctly predicts whether a customer will default or not 81.20% of the time. This means that out of every 100 customers in the test set, the model makes correct predictions for about 81 of them. The narrow confidence interval shows that this result is fairly reliable and wouldn't change much if we tested on different data.

Sensitivity (24.72%): Default Detection Rate - This is actually one of the challenging aspects of our model. The model only catches about 24.72% of customers who actually default. In practical terms, if there are 100 customers who will truly default, our model would identify only about 25 of them. This means we miss 75% of the actual defaults (false negatives). This low sensitivity happens because our model is being conservative - it's trying hard not to falsely alarm customers who won't default.

Specificity (97.23%): Non-Default Identification - On the flip side, our model is very good at identifying customers who will NOT default. It correctly identifies 97.23% of non-defaulters. This is actually a strength because it means we rarely bother customers who are good payers with unnecessary interventions or warnings.

Precision (71.72%): Prediction Reliability - When our model predicts that someone will default, it's correct about 71.72% of the time. This means if we act on the model's default predictions, we can be fairly confident that most of those customers are actually at risk.

Balanced Accuracy (60.98%) - Since our dataset is imbalanced (22% defaults, 78% non-defaults), we use balanced accuracy to get a fair picture. This gives equal weight to sensitivity and specificity, showing that on average, our model performs moderately well on both classes.

AUC (0.7250): Model Discrimination Ability - This tells us how well our model can distinguish between defaulters and non-defaulters at different probability thresholds. A score of 0.72 is reasonably good (better than 0.50 which is random), though there's definitely room for improvement toward 0.80 or higher.

What the Trade-offs Mean

Notice something interesting: our model is very cautious. It's fantastic at identifying customers who are fine (97% specificity) but misses quite a few who actually default (only 25% sensitivity). Why? Because:

1. Most people don't default, so the model learns that "non-default" is the safer prediction
2. The 0.5 cutoff (we say "default" if probability > 0.5) was kind of arbitrary - we could change it
3. Our training data naturally has this imbalance, so the model reflects it

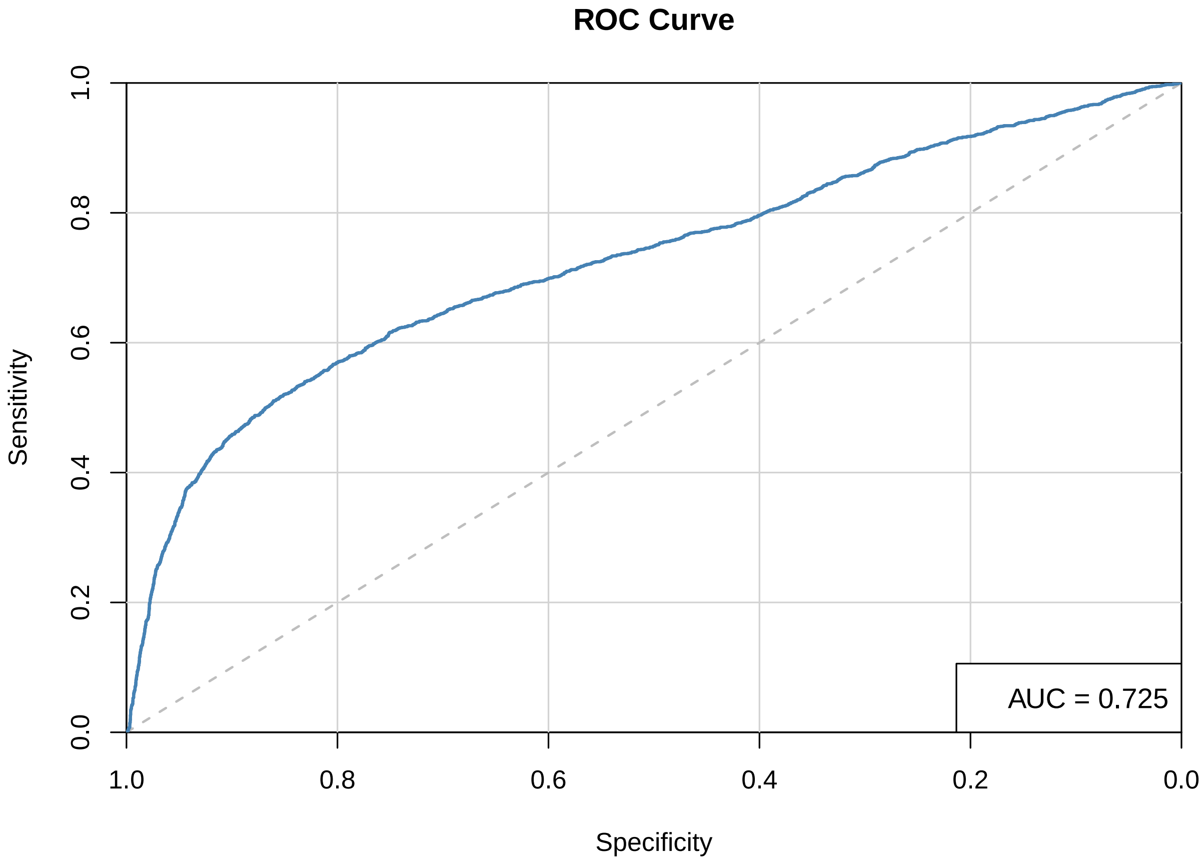
In the real world, this kind of model would probably be used to flag high-risk customers for extra attention, not as the only tool for making decisions about who gets credit.

Practical Uses and Limitations

- **Good at**: Not bothering customers who are fine (97% of non-defaulters get flagged as safe)
- **Not great at**: Catching all the risky customers (we'd miss 75% of actual defaults)
- **Works best for**: Obvious risk cases like recent late payments or very old accounts
- **Might miss**: Defaults by customers who seemed fine until suddenly they weren't
- **Could be improved**: If we adjust the threshold to be more aggressive, we'd catch more defaults, but we'd also bother more good customers

ROC Curve and AUC

The ROC curve shows us how good our model is at distinguishing between defaulters and non-defaulters, across all possible decision thresholds.



Understanding Our AUC Score

AUC: 0.7250

AUC tells us how well the model can separate defaulters from non-defaulters..

The ROC curve itself shows the different trade-offs we face. Right now we're using 0.5 as the cutoff (predict default if probability > 0.5). But we could slide that threshold left or right depending on what matters more to us: catching defaults or avoiding false alarms. The curve shows all those options.

Conclusion

Building a predictive model for credit card defaults is not a simple task, and our model reflects this complexity. The results show that we have developed a reasonably effective tool, though it comes with both strengths and limitations that must be understood before implementation.

On the positive side, our model achieves 81% overall accuracy, correctly classifying most customers in our test set. When the model predicts that a customer is safe, it is correct approximately 97% of the time. This high specificity means that the model rarely triggers false alarms for good customers, which is valuable for maintaining customer relationships. Additionally, when the model does flag someone as being at risk of default, it is correct about 72% of the time, making these predictions meaningful enough to warrant further investigation. Perhaps most importantly, the risk factors our model identified align well with intuition: recent payment delays indicate higher risk, while factors like marital stability and higher education levels are protective.

However, there are significant limitations we must acknowledge. The model only catches about 25% of customers who actually default, which means we miss three-quarters of the people who will truly default. This happens because most customers don't default, so the model learns to be very conservative in its predictions, favoring the safer assumption. The AUC of 0.72 indicates decent discriminatory ability but leaves room for improvement. These limitations suggest that relying solely on this model for credit decisions would be insufficient.

The practical application of this model is therefore best understood as part of a larger risk management system. Rather than using it as the sole decision-making tool, the model works best as a screening mechanism to identify the highest-confidence default cases for closer manual review. Banks might use these predictions to flag customers for additional scrutiny, to combine with other assessment methods, or to adjust their lending strategies accordingly. The trade-off is clear: if we wanted to catch more defaults, we could lower the decision threshold, but this would inevitably increase false alarms for good customers.

In all, our analysis confirms what many financial institutions already understand: payment history is the strongest predictor of default risk. Recent payment behavior, more than any other factor, reveals a customer's likelihood to default. Our model provides empirical support for this intuition and demonstrates that while predicting defaults is challenging, it is possible to build models that add value to decision-making processes when used appropriately within a broader framework.