# Exploration of Linear Classifiers: Predicting Repayment Status on Credit Card Dataset

Dayu Zhong, Yifei Sun, Zhangzhi Cao, Yue Shi

## Abstract

The report applies supervised learning models including logistic regression, support vector machine and random forest on a credit card dataset to predict whether customers will default or not next month (October 2005) based on their credit history in the last 6 months (from April to September 2005), as well as personal background such as marriage status and education. The report, in particular, utilizes the panel data approach to explore the relationship among amount of previous payment, repayment status and amount of bill statement.

## 1. Introduction

Back in 2005, credit card issuers in Taiwan faced a cash and credit card debt crisis, with delinquency expected to peak in the third quarter of 2006. In order to increase market share, card-issuing banks in Taiwan over-issued cash and credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit cards for consumption purposes and accumulated heavy cash and credit card debts. This crisis caused a blow to consumer financial confidence and presented a big challenge for both banks and cardholders.

For banks, there is a balance between reward and risk, credit card issuing-banks can earn money from annual fees, interest earned on outstanding balance of credit card holders; on the other hand, they also face a risk of default payment of credit card clients.

Banks prefer two type of customers. The first type is the ones that never default but would pay up the minimum amount so they would charge interests on credit balance. The second types is the ones that always miss the minimum payments but can pay up later so that the banks can charge them additional late fees. It is ambitious to study how banks can maximize their profits by making decisions on credit issuance since banks might prefer customers that defaults from time to time. This would require a thorough analysis of the personal and credit background of the clients.

On the other hand, the bank might be interested in learning about whether a customer will default or not based on his profile. We are looking at the credit crisis of 2005 in Taiwan. It might be possible that banks are very cautious of customers that could potentially default. Suppose a bank in September 2005 asks us to predict which customers would default next month in October 2005, we are provided with the following dataset, which consists of default payments from a major credit card company in Taiwan 2005. The data set consists of 13,276 instances and 24 attributes consisting of gender, education profile, marital status, age, history of statement balance, payment status and binary status of default (1 or 0).

- LIMIT_BAL: Amount of given credit limit in NT dollars (includes individual and family/supplementary credit
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_1: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- Default.Payment.Next.Month: (October, 2005) Default payment (1=yes, 0=no)

In this report, we are interested in predicting Default_Payment_Next Month, i.e., whether a customer will default or not in October 2005. We in particular want to learn in depth about linear classifiers such as logistic regression, support vector machine and random forest and evaluate their performance on this dataset. We conduct exploratory data analysis to understand the dataset at first.

## 2. Exploratory Data Analysis

We first explore the demographical variables in our dataset. Figure 2.1 shows the gender distribution of credit cardholders. For this variable, 1 means male and 2 means female. It is clear see that there are more female clients than male clients. Figure 2.2 shows the cardholder's education (1=graduate school, 2=university, 3=high school, 4=others). It is obvious that the number of clients who have university-level education is the largest, followed by clients having graduate-level education. It seems that higher education receivers are more likely to have a credit card. Figure 2.3 reveals that the number of credit cardholders has no apparent relationship to marital status. There are more single clients than married, but the number is quite close. Figure 2.4 displays a right-skewed distribution for age, with the peak around 28 years old. This condition may be caused by the phenomenon that credit cards are more acceptable and popular among young people.

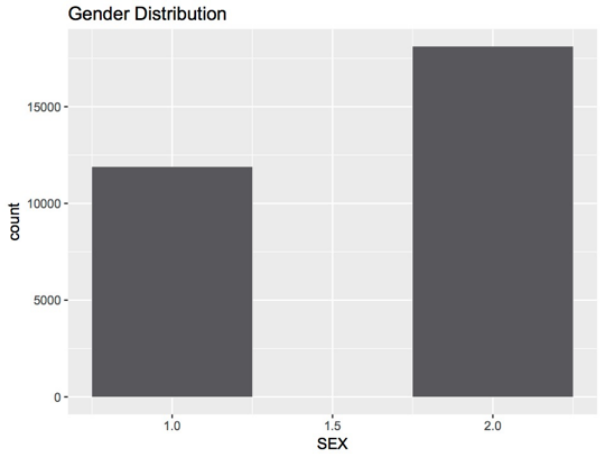



Figure 2.1

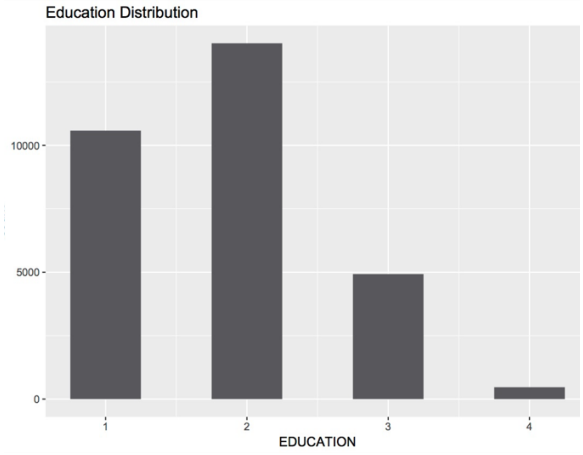



Figure 2.2

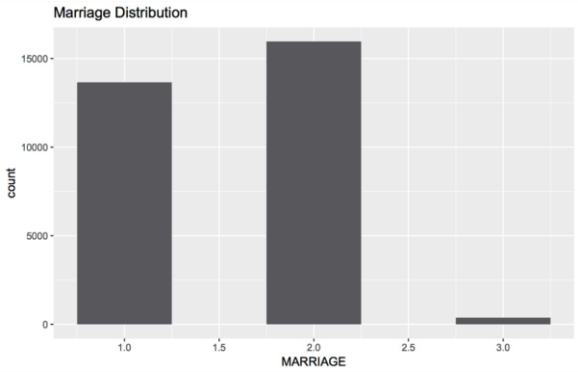



Figure 2.3

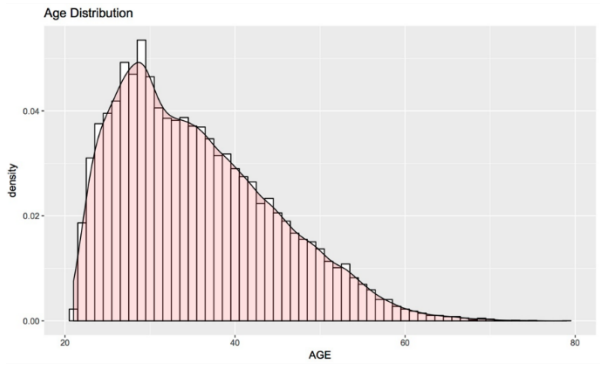



Figure 2.4

Now we move onto the credit history of clients. At first we consider LIMIT_BAL, which is the maximum amount a person is allowed to borrow on a credit card. It includes purchases, cash advances, and any finance charges or fees. From the density plot (Figure 2.5), we could see that the majority of the credit limit is between NT$60,000 and NT$400,000.

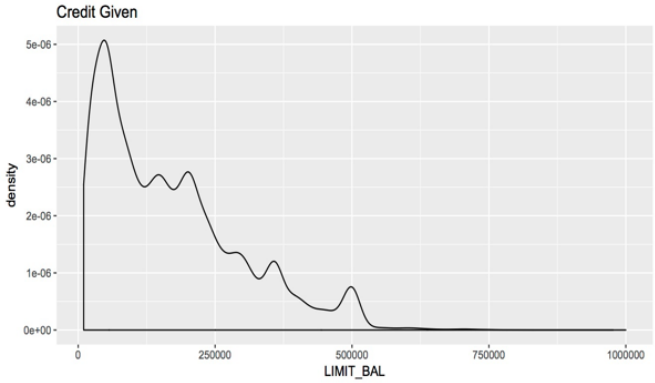


Figure 2.5

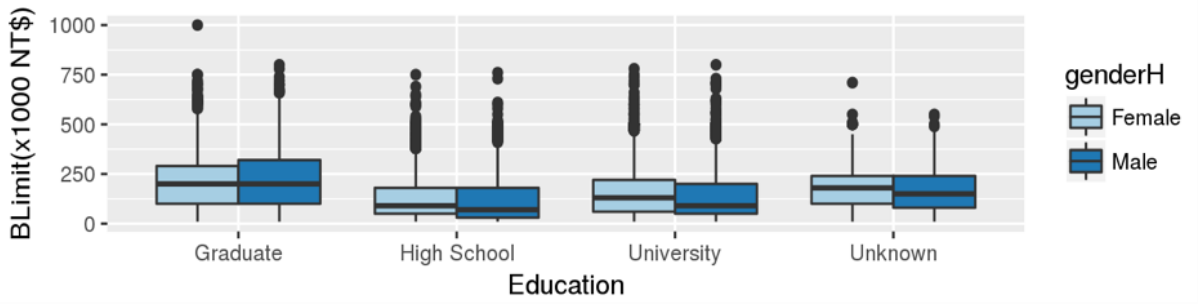
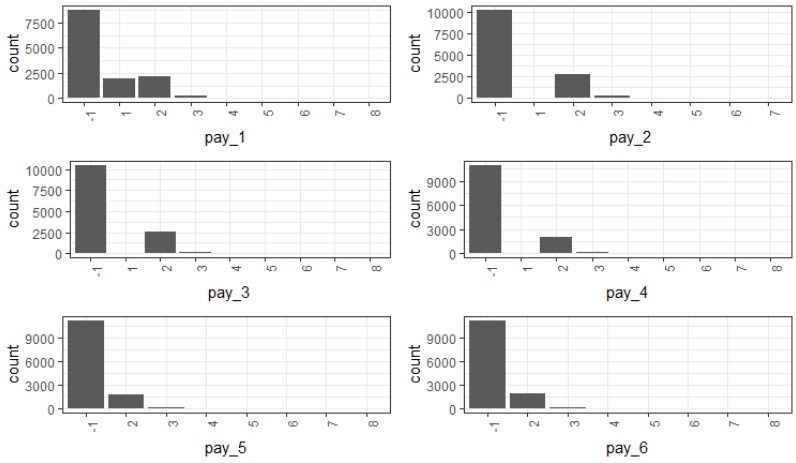



Figure 2.6

Figure 2.6 suggests that college students and graduate students tend to have a higher credit limits than those who only attended high school. This makes sense intuitively.




-1=pay duly,
1=payment delay for one month,
2=payment delay for two months,
3=payment delay for three month,
4=payment delay for four month,
5=payment delay for five month,
6=payment delay for six month,

Figure 2.7

Figure 2.7 shows the repayment status from April to September. Around 81% credit card holders pay the bill on time, and we observe as time goes by, there is a trend that most people pay delay for one month, two months, three months. Literally no one pay delay more than four month. If the bank want to earn interest on credit card holders, it may mostly consider on the holders who default for two months.

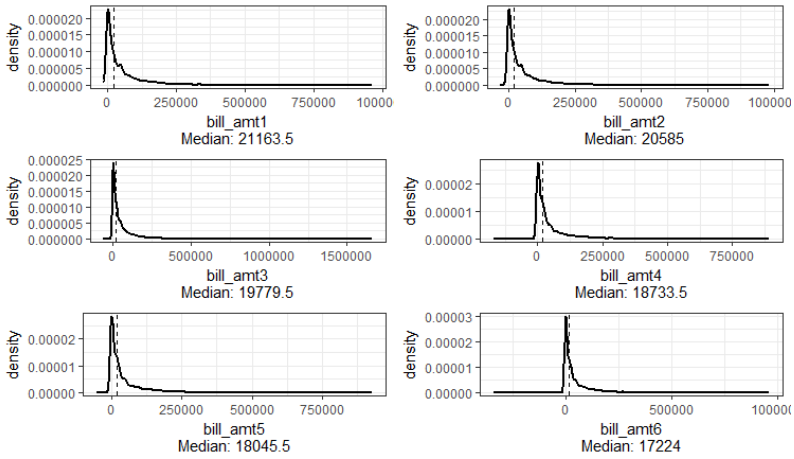



Figure 2.8

Figure 2.8 shows the density distribution of Bill Amount from April to September. Since the distribution is skewed to the right, we use median to show the average, there is a trend that average bill amount increases each month. This is because the accumulative delay increases during the six month as we showed in last graph.

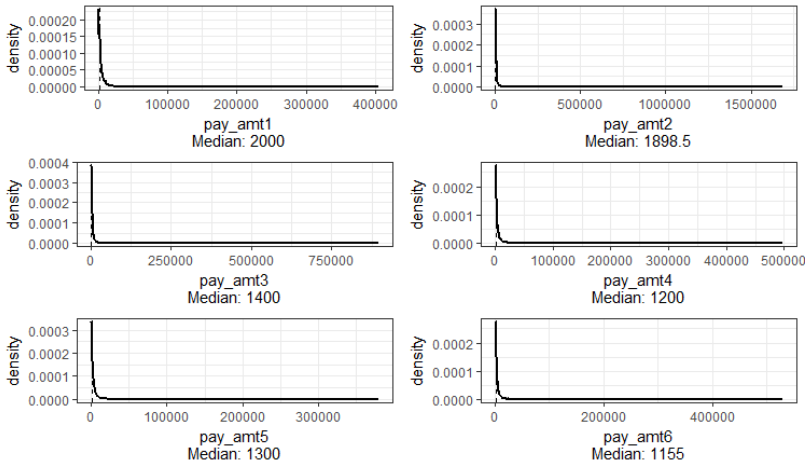



Figure 2.9

Figure 2.9 shows the density distribution of Payment Amount from April to September. Also use the median to show the average, and there is a trend that is average payment amount increases each month. This is because as the bill payment increase, the minimum payment for each month increases, as a result credit card holders have to pay more.

Finally we look at the default payment status next month in our dataset, our response variable. Defaulting vs not defaulting is roughly a 50/50 split.

## 3. Model Building

### 3.1 Logistic Regression

We want to predict whether a customer will default or not next month based on his personal background such as age and education as well as his credit history, which includes default history and bank statement balance in the last 6 months. We first develop logistic regression model to solve this classification problem.

The logistic regression assumes that all the points/samples in the dataset can be separated by a smooth linear boundary. It compares the probability of default versus not default, 1 versus 0 in our case. It will predict that a customer will default next month if the trained model suggests that the probability of defaulting is higher than the probability of not defaulting.

$$\frac{P(y=1|x)}{P(y=0|x)} \geq 1 \rightarrow \log\frac{P(y=1|x)}{P(y=0|x)} \geq 0 \quad = w_0 + w * x$$

$$Since\ P(y=1|x) = 1 - P(y=0|x), we\ can\ derive\ the\ following\ decision\ boundary.$$

$$P(y=1|x) = \frac{1}{1+\exp(-w_0 - w*x)} = \frac{1}{2} \quad \rightarrow \quad w_0 + w*x = 0$$

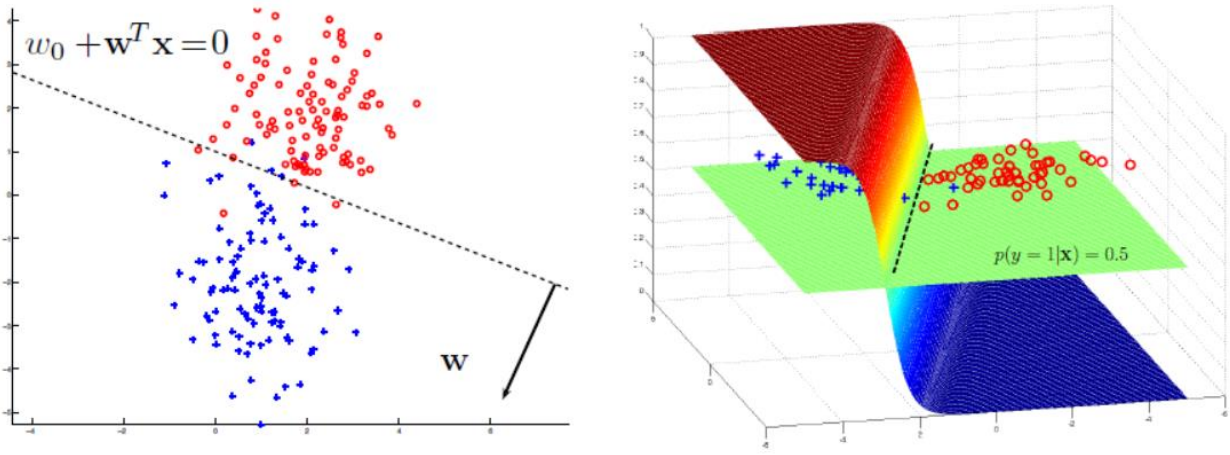



Figure 3.1

Figure 3.1 visualizes the decision boundary for logistic regression respectively for two and three predictor variables. Because we have more than 20 predictor variables in our model, we could not visualize the boundary directly but we can observe the model coefficients.

We randomly shuffle the dataset into 70/30 split as training set and testing set. Below is our trained logistic model.

```
glm(formula = default.payment.next.month ~ ., family = binomial,
    data = train.batch)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2637  -0.9563  -0.4186   1.0563   2.9258

Coefficients:
                 Estimate    Std. Error z value      Pr(>|z|)
(Intercept) -0.2415373733  0.1845924130  -1.308      0.190707
limit_bal   -0.0000012697  0.0000002210  -5.746 0.00000000913 ***
sex         -0.0151932616  0.0472205965  -0.322      0.747642
education    0.0044847126  0.0336207708   0.133      0.893884
marriage    -0.1903856100  0.0486875256  -3.910 0.00009215978 ***
age          0.0065031518  0.0027832753   2.337      0.019465 *
pay_1        0.8080303835  0.0397813278  20.312       < 2e-16 ***
pay_2        0.0925226191  0.0408171369   2.267      0.023405 *
pay_3        0.1325011580  0.0429067874   3.088      0.002014 **
pay_4        0.2004948376  0.0494026950   4.058 0.00004941462 ***
pay_5        0.0598192333  0.0522197481   1.146      0.251990
pay_6        0.0772421041  0.0442802424   1.744      0.081091 .
bill_amt1   -0.0000014596  0.0000014402  -1.013      0.310857
bill_amt2   -0.0000004217  0.0000019979  -0.211      0.832837
bill_amt3    0.0000011431  0.0000018861   0.606      0.544453
bill_amt4    0.0000068432  0.0000019055   3.591      0.000329 ***
bill_amt5   -0.0000027766  0.0000021359  -1.300      0.193617
bill_amt6   -0.0000023017  0.0000016420  -1.402      0.160982
pay_amt1    -0.0000121788  0.0000031018  -3.926 0.00008622954 ***
pay_amt2    -0.0000086552  0.0000024005  -3.606      0.000312 ***
pay_amt3    -0.0000035155  0.0000025194  -1.395      0.162902
pay_amt4     0.0000002738  0.0000024505   0.112      0.911030
pay_amt5    -0.0000041025  0.0000026319  -1.559      0.119061
pay_amt6    -0.0000011348  0.0000017466  -0.650      0.515880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12878  on 9289  degrees of freedom
Residual deviance: 10946  on 9266  degrees of freedom
AIC: 10994
```

Figure 3.2

We first evaluate the fitness of the logistic regression. The null deviance is very large compared to the degree of freedom, suggesting that the null model, which includes only the intercept term, does not fit the dataset well. The residual deviance, which corresponds to the full model, is 10946 under 9226 degrees of freedom. It has relatively big improvement over the null model but the high deviance still suggests that the logistic regression may not be a very good fit for the dataset.

Now we look at the coefficients of the logistic regression. While most of the coefficients for the variables have low p values, sex and education along with bill_amt1, bill_amt2, bill_amt3, pay_amt4, and pay_amt6 do not appear to be statistically significant. We note that while some variables have very small coefficients, they might have significantly large values to compensate.

We apply the trained model on the testing dataset and achieves an accuracy of 66.98%. This suggests that our logistic regression performs better than the naïve model where we would predict that every single customer will default next month. Since we have a rough 50/50 split between

defaulting and not defaulting in our dataset, we would achieve 50% accuracy under the naïve model.

While the logistic regression has achieve high accuracy of 66.98%, we do not want to stop here and be satisfied. We want to learn about places where logistic regressions made an error and see if we can either improve on the current model or find another model that addresses the flaws.

We plot the confusion matrix of our logistic model to break down its performance.

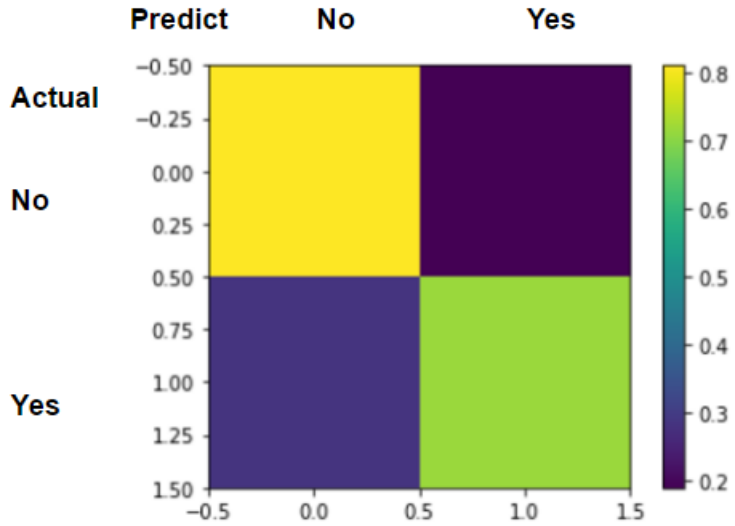



Figure 3.3

The upper left square in figure 3.3 is filled with color yellow, which correspond to a ratio of 0.8. This means that the logistic model correctly classifies not defaulting 80% of the time while wrongly classifying the defaulting as not defaulting 20% of the time, which corresponds to the lower left square. Likewise, the lower right square, with color green, suggests that the logistic model only correctly classifies defaulting 60% of the time and therefore make mistakes of wrongly predicting not defaulting as defaulting 40% of the time.

We see that the logistic regression has a high false positive rate. We plot the default score/probability distribution to investigate the reason that the logistic regression tends to make type I error.

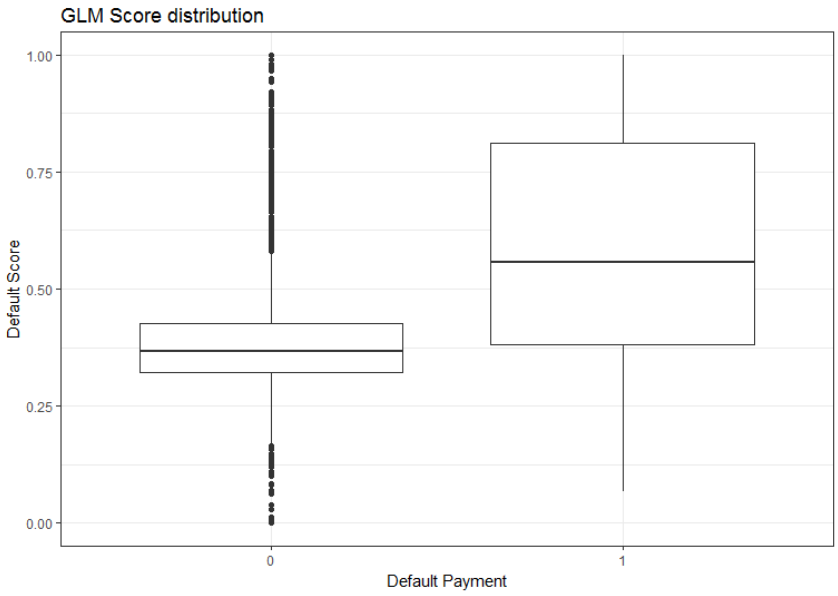



Figure 3.4

Figure 3.4 shows the probability distribution of our dataset under the logistic model. On the left side of the plot, we see that the 95% confidence interval of default probability when a customer does not default is between 0.32 and 0.45, which is smaller than 0.5. This suggests that the logistic model does a good job of predicting not defaulting. On the right side of the plot, the 95% confidence interval of default probability is between 0.375 and 0.8, which is not always greater than 0.5. Logistic regression would wrongly predict those customers to default when they in reality do not. Our interpretation of the probability distribution is in accordance with that of the confusion matrix, namely that logistic model tends to wrongly classify not defaulting as defaulting.

One important observation that we want to emphasize is that the dataset seems to cluster around probability/score of 0.5 under the logistic model. That is, there is a high portion of dataset with probability ranging from 0.4 to 0.6. There are many points close to the decision boundary where the probability is equal to 0.5. Logistic regression would tend to make mistakes in this case because the linear classifier only consider the likelihood of the data points. It does not consider the distance of points to the decision boundary.

Consider the following plot, figure 3.5.

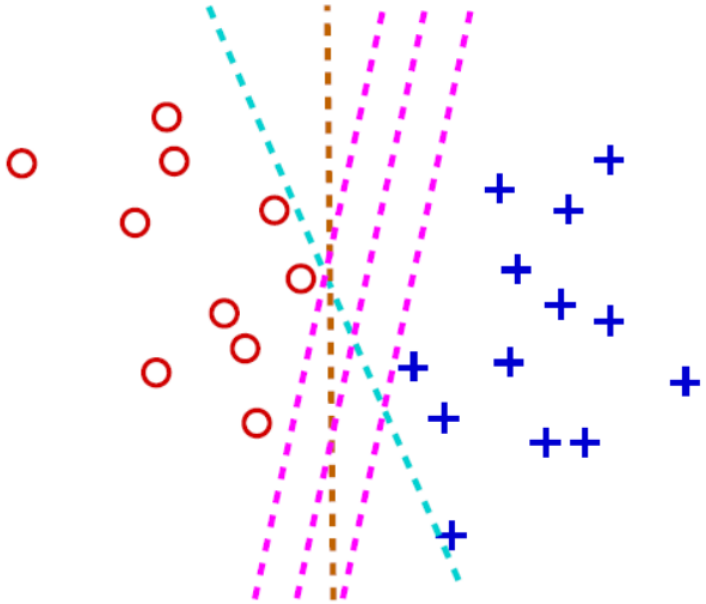


Figure 3.5

The blue plot is the decision boundary produced by the logistic regression. As we add more outliers of class "+" in the dataset, the decision boundary changes to brown and then the left most pink line. The original blue separately the data well, but the risk of misclassifying is high since the points are very close to decision boundary. As we add outliers to the dataset, the direction of decision boundary changes dramatically. While it is true that the logistic regression uses the log loss function and therefore is less subject to the influence of outliers, it will be sensitive to their influence in our case because a high potion of our dataset cluster around probability 0.5 under the logistic model.

We believe that the second pink line (the one in the center) seems to be the best linear classifier. And here we motivate the use of support vector machine. The support vector machine model would for sure select the second pink line as the linear decision boundary as it is a max margin algorithm that maximizes the distance to the closest points. When adding more outliers of class "+", the decision boundary remain the same under the support vector machine model.

### 3.2 Support Vector Machine

We are interested in finding a linear classifier that takes into consideration distance to the nearby points. We want to find a large margin classifier that minimizes the distance to the closest points: $argmax\{\frac{1}{|w|}\min_i y_i(w * x_i + w_0)\}$. Without loss of generality, we can assume that $\min_i y_i(w * x_i + w_0) = 1$. Now our optimization problem becomes

$$argmax\left\{\frac{1}{|w|}\right\} s.t. y_i(w * x_i + w_0) \geq 1 \rightarrow \ \ argmin\{|w|^2|\} \, s.t. y_i(w * x_i + w_0) \geq 1$$

In logistic regression, the loss function is the negative log loss. For the support vector machine, it uses the hinge loss function defined below:

$$\max_{\alpha_i} \alpha_i[1 - y_i(w * x_i + w_0)] = 0 \;\; if \; y_i(w * x_i + w_0) \; - 1 \geq 0; = infinity \; otherwise$$

Figure 3.6 compares the log loss (log p) with hinge loss.

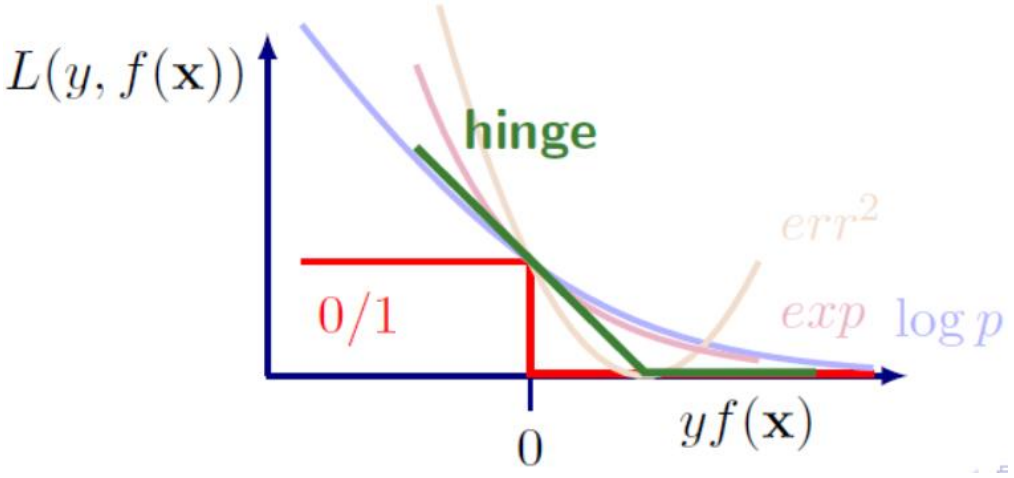



Figure 3.6

We can reformulate our problem now: $\min_w \frac{1}{2}|w|^2 + \sum_i \max_{\alpha_i \geq 0} \alpha_i[1 - y_i(w * x_i + w_0)]$

$$= \min_w \max_{\alpha \geq 0} \{\frac{1}{2}|w|^2 + \sum_i \alpha_i[1 - y_i(w * x_i + w_0]$$

$$= \max_{\alpha \geq 0} \min_w \frac{1}{2}|w|^2 + \sum_i \alpha_i[1 - y_i(w * x_i + w_0] = \max_{\alpha_i}\{J\}$$

The above equality holds because the duality constraints for support vector machine were satisfied under the KKT conditions. We can then proceed to find the close form for $w$ by taking the derivative of J with respect to $w \; and \; w_0$: $w - \sum_i \alpha_i y_i x_i = 0 \quad -\sum_i \alpha_i y_i = 0$

With some algebra, the problem becomes $\max_{\alpha \geq 0, \sum_i \alpha_i y_i = 0} \{\sum_i \alpha_i \; - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i * x_j\}$

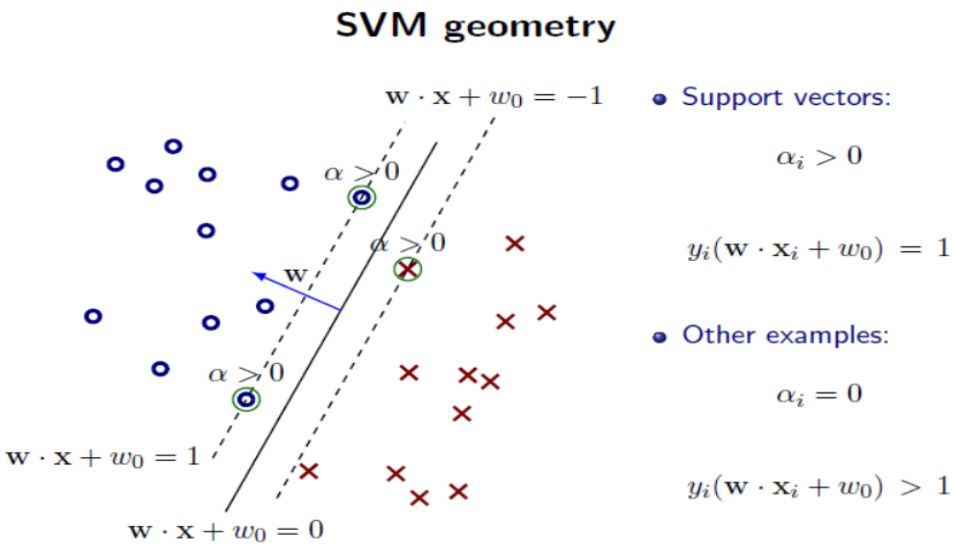



Figure 3.7

Figure 3.7 above shows an example of decision boundary by support vector machine. Note that this assumes this data is linearly separable.

When this assumption does not hold, we have to introduce a slack variable to penalize misclassification, which can be shown to be equivalent to adding a constraint to alpha. The resulting dual problem becomes $\max_{\sum_i \alpha_i y_i = 0,\ 0 \le \alpha \le C} \{\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i * x_j\}$

We can use quadratic programming algorithm such as SMO to solve this problem. But in our case, we apply cvx.opt package in python 2.7 to solve it. In applying the support vector machine model, we convert categorical variable into one-hot encoding, ordinal variables into unary representation and also normalize continuous variables. We preprocess the data so that the effective variation in the continuous feature value should not be much greater than one. Otherwise the margin of 1 may be dominated by a single feature. After we trained the support vector machine on the training dataset, we applied the model on our test set and achieved 76.49% accuracy under the linear kernel. Recall the logistic regression has 66.98% accuracy.

Before we hold the belief that the support vector machine would perform better than logistic regression for this dataset and in particular improve the classification issues around the decision boundaries, therefore reducing the false positive rate. We compare the confusion matrices and see if this is true.

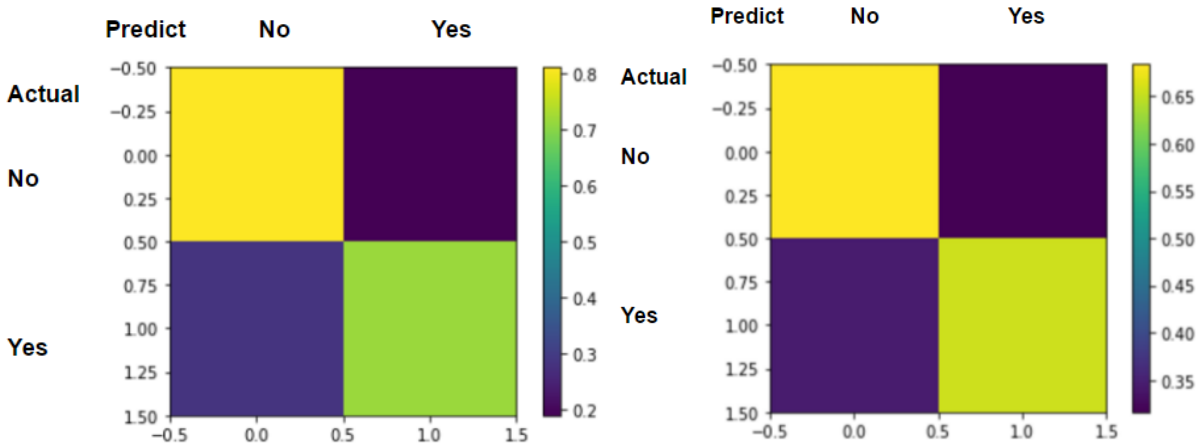



Figure 3.8 Logistic Regression Support Vector Machine

Compared with logistic regression, support vector machine has a higher true positive rate (sensitivity) and true negative rate (specificity). Indeed, the false negative rate decreases by roughly 10%, which is a significant improvement.

### 3.3 Regression trees

So far we have considered two different types of linear classifier: logistic regression and support vector machine (although for support vector machine we can also apply other nonlinear kernels). We now move onto regression trees. Why? Because Regression trees are closely related to the two models we have developed. Trees are simply stepwise linear functions. It consists many constant lines that are parallel to axes. Suppose we have a tree with M leaves, let leaf m correspond to region $R_m$, $f(x) = \sum_m f_m[\, x \in R_m] \; where \; [A] = 1 \; if \; A \; and \; 0 \; otherwise.$

For our classification problem, we want to minimize 0/1 loss, i.e. $argmin_{y\hat{}_m} \sum_m \left[ y_i \neq \hat{y_m} \right]$. And to minimize this loss, regression trees apply a top-down greedy approach called recursive binary splitting. At each split, it selects a predictor $X_i$ and a cut point $s$ such that splitting the predictor space into regions $\{X|X_i<s\}$ and $\{X|X_i \geq s\}$ leads to the greatest possible reduction for the 0/1 loss function.

We can grow a tree into as many as branches as the number of points in our dataset and thus achieve a training loss of 0. But such a tree would definitely overfit the dataset, leading to a high validation error. We therefore have to prune the tree by collapsing some of the internal nodes.

We first grow a large tree $T_0$ and then prune to $T \in T_0$. We introduce some notations here.

$$|T| \; number \; of \; leafs \; in \; T \; ; N_m = |I_m| = \{i : x_i \in R_m\}; \; f_m the \; value \; in \; the \; leaf;$$

$$Q_m(T) = \sum_c \hat{p_{m,c}} * \left(1 - \hat{p_{m,c}}\right) \; where \; \hat{p_{m,c}} = \frac{1}{N_m} \sum_{i \in I_m} [y_i = c]$$

The cost-complexity criterion (CART) of tree $T \in T_0$ is $C_\lambda(T) = \sum_{m=1}^{|T|} N_m Q_M(T) + \lambda T$. T is obtained from $T_0$ by merging multiple leafs. We keep collapsing the internal nodes that produces

the smallest increase in $\sum_m N_m Q_m(T)$, going from $T_0$ to a single node. Then for a given $\lambda$, there exists a unique $T_\lambda = argmin_T\, C_\lambda(T)$. Below shows the pruning process of the first two nodes.

```
          CP nsplit rel error    xerror        xstd
1 0.35590653      0 1.0000000 1.0144959 0.010426257
2 0.02553007      1 0.6440935 0.6440935 0.009731270
3 0.01000000      2 0.6185634 0.6185634 0.009625173

Variable importance
   PAY_1    PAY_2    PAY_4    PAY_3    PAY_5 PAY_AMT1
      37       21       13       12        9        8

Node number 1: 9290 observations,    complexity param=0.3559065
  predicted class=0  expected loss=0.4975242  P(node) =1
    class counts:  4668  4622
   probabilities: 0.502 0.498
  left son=2 (6163 obs) right son=3 (3127 obs)
  Primary splits:
      PAY_1 < 0      to the left,  improve=664.5609, (0 missing)
      PAY_2 < 1.5    to the left,  improve=534.3059, (0 missing)
      PAY_3 < 1.5    to the left,  improve=403.4042, (0 missing)
      PAY_4 < 1.5    to the left,  improve=390.2155, (0 missing)
      PAY_5 < 0.5    to the left,  improve=315.5177, (0 missing)
  Surrogate splits:
      PAY_2    < 0      to the left,  agree=0.856, adj=0.572, (0 split)
      PAY_3    < 0      to the left,  agree=0.770, adj=0.315, (0 split)
      PAY_4    < 0      to the left,  agree=0.750, adj=0.258, (0 split)
      PAY_5    < 0.5    to the left,  agree=0.742, adj=0.233, (0 split)
      PAY_AMT1 < 76.5   to the right, agree=0.738, adj=0.223, (0 split)

Node number 2: 6163 observations,    complexity param=0.02553007
  predicted class=0  expected loss=0.3628103  P(node) =0.6634015
    class counts:  3927  2236
   probabilities: 0.637 0.363
  left son=4 (5765 obs) right son=5 (398 obs)
  Primary splits:
      PAY_4    < 0.5    to the left,  improve=69.32787, (0 missing)
      PAY_AMT1 < 2000.5 to the right, improve=55.21570, (0 missing)
      PAY_AMT5 < 1101.5 to the right, improve=55.14642, (0 missing)
      PAY_3    < 0.5    to the left,  improve=53.47328, (0 missing)
      PAY_AMT4 < 1000.5 to the right, improve=52.81179, (0 missing)
  Surrogate splits:
      PAY_5 < 0.5    to the left,  agree=0.943, adj=0.123, (0 split)
      PAY_3 < 2.5    to the left,  agree=0.939, adj=0.048, (0 split)
      PAY_6 < 2.5    to the left,  agree=0.936, adj=0.010, (0 split)
      PAY_2 < 3.5    to the left,  agree=0.936, adj=0.008, (0 split)
```

Figure 3.9

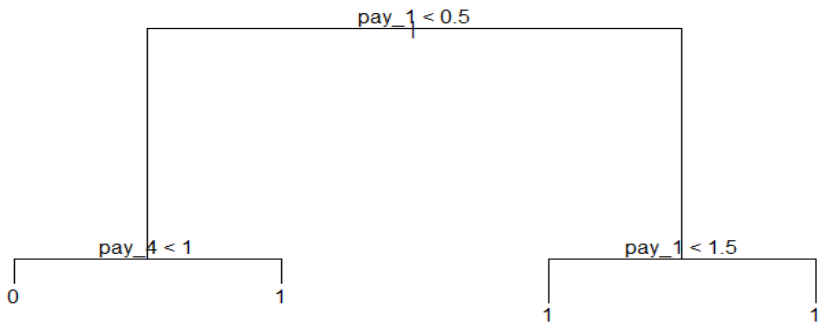



Figure 3.10

Figure 3.10 shows the pruned tree $T_\lambda$ with the smallest cost criterion but not necessarily the best tree.

We might have over-pruned the trees in this case, thus under fitting the dataset. Here a small change of threshold in pay_1 and pay_4 would drastically change the result. We therefore combine multiple regression trees and utilize the method of bagging.

Bagging can be used to reduce the variance of regression trees. We can construct regression trees using bootstrapped training sets and average those predictions. Each tree has high variance but low bias. Applying the central limit theorem, we take the vote of these trees to reduce the total variance.

Random forest improves upon the procedure of bagging by reducing the correlations between the regression trees. It builds T trees independently (in parallel). For each tree, sample N points with replacement and grow a CART tree where in each node it only look at a subset of m = $\sqrt{d} < d$ features where the trees are not pruned. Our trained model of random forest is shown below, which achieves 85.08% accuracy on test set. Recall the logistic regression has 66.98% accuracy and the support vector machine has 76.49% accuracy. The random forest we built from averaging the vote from multiple stepwise linear functions, has yield more than 18% improvement on accuracy over logistic regression and 8.5% improvement over support vector machine.

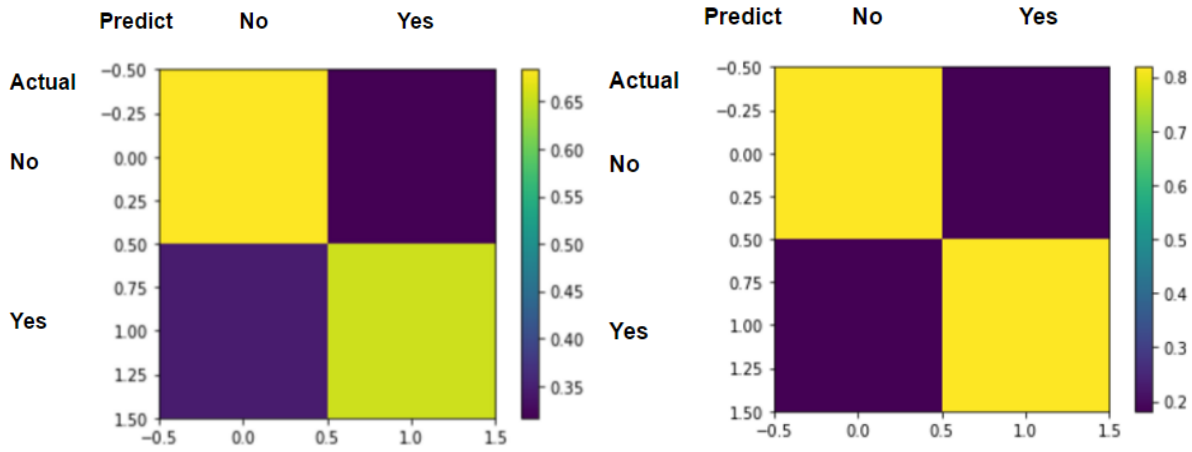



Figure 3.11 Support Vector Machine Random Forest

The confusion matrix suggests that random forest has both has a higher true positive rate (sensitivity) and true negative rate (specificity) than support vector machine, and therefore seems to be the best model among the three “linear” models that we have selected

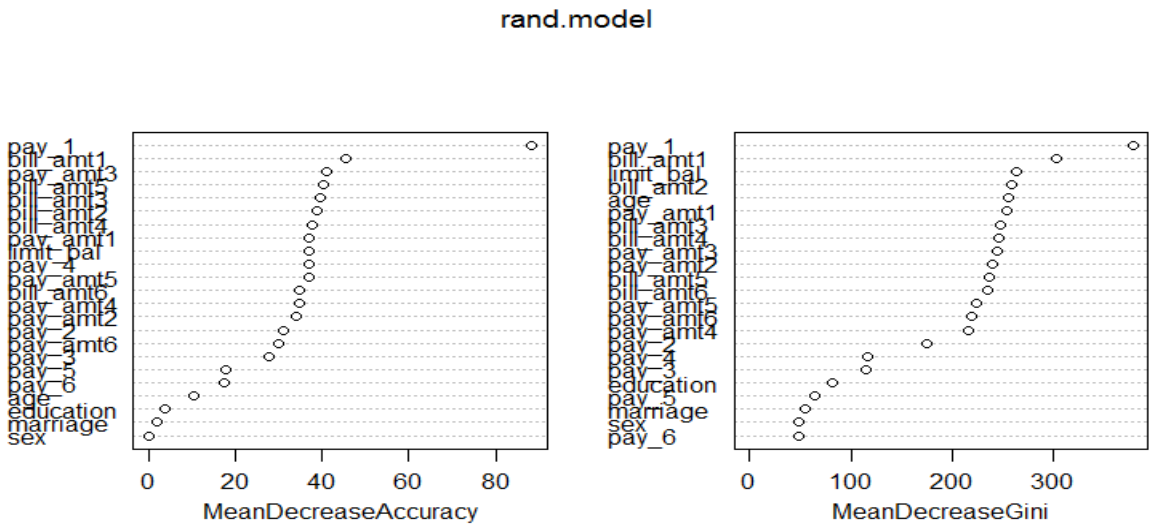



Figure 3.12

The mean decreased accuracy measures how much more helpful than random a particular predictor variable is in successfully classifying data. Similarly, the mean decreased Gini index(decrease in gini index) suggests the importance of role a predictor variables plays in partitioning the dataset as we would prefer a gini index.. From the plot Figure 3.12, it appears that repayment status in August pay_1, the amount of bill statement in September bill_amt1 and the amount of previous payment in September pay_amt3 are the most important. This result makes perfect sense since we are predicting whether the customer will default or not in October. The most recent past credit history should be the most indicative of the next payment status.

And overall, we can see that the repayment status, amount of bill statement, amount of previous payment are the most important variables. This is consistent with the results of the logistic regression. Figure 3.13 suggests that the amount of previous payments pay_amt_i have very low correlation among one another, whereas the repayment status pay_i, as well as previous bill statement bill_amt_i, are highly correlated with one another.

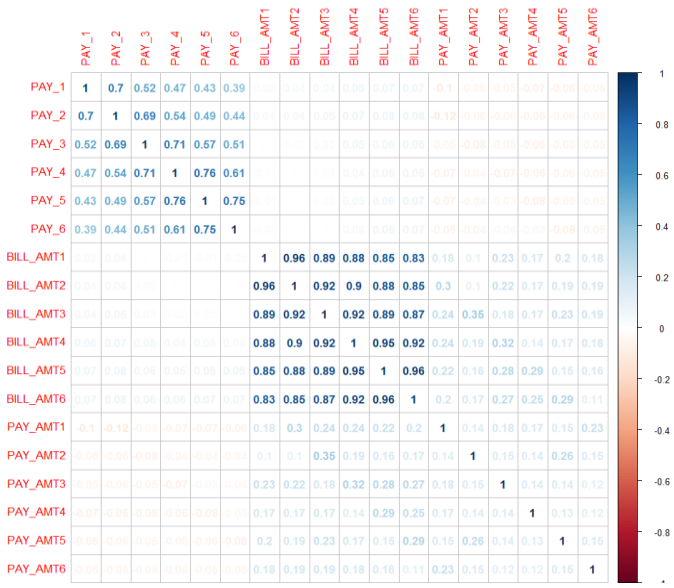



Figure 3.13

Now we want to further explore these 18 variables by building a panel data model.

### 3.4 Time Series Analysis?

We first wanted to build a multivariate time series model. The *vector autoregression (VAR) model* seems to be a good fit. However, like normal time series analysis, there is requirement for sample size. *Hanke* and *Wichern* recommend a minimum $2 \times s$ to $6 \times s$ depending on the method (where s is the seasonal period, so s=12 for monthly data). We couldn't find out a seasonal pattern since we only have data for half of a year.

Besides, we do not see an obvious pattern as time goes. We randomly pick 9 observations of amount of previous payment and amount of bill statement respectively and draw the following plot Figure 3.14.

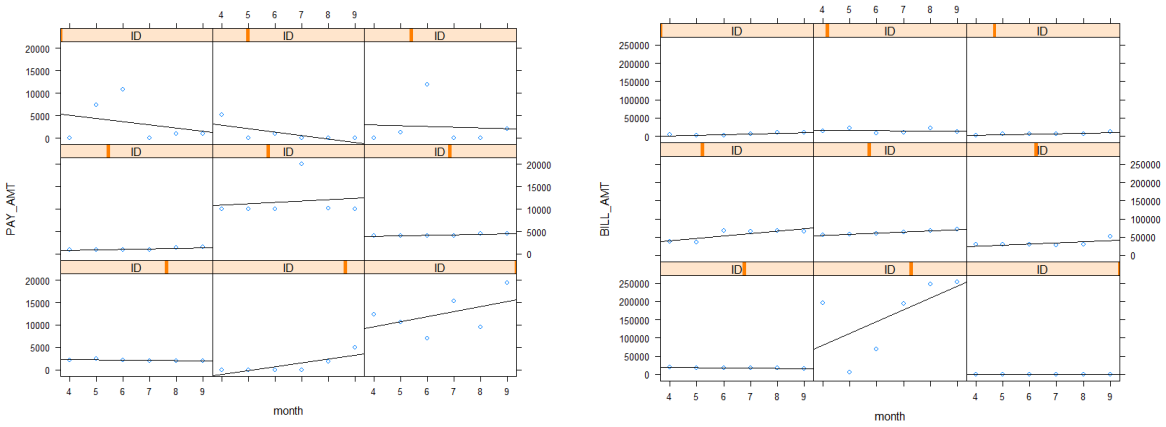



Figure 3.14

Amount of previous payment increases or decreases from April to September. For some customers, the amount stays relatively stable. Likewise, there is no obvious variation trend for amount of bill statement.

## 3.5 Panel Data Analysis

### 3.5.1 Introduction

Now we apply the panel data approach to understand the relationship between repayment status, amount of bill statement and amount of previous payment. In particular, we are interested in solving the following problem Panel data analysis represents a marriage of regression and time series analysis. Panel data involve two dimensions: a cross-sectional dimension N, and a time-series dimension T. As with many regression data sets, panel data are composed of a cross-section of subjects. However, unlike regression data, with panel data we observe subjects over time. At the same time, unlike time series data, with panel data we observe many subjects.

Traditional time series are panel data observed over a long time period on a single experimental unit. In our dataset, we only have data for six months given each customer. Thus, traditional time series could not be properly applied. Here panel data analysis gives us a good solution. Panel data time series usually have fewer repeated measurements than traditional time series data. The average number of observations per subject will commonly be anywhere from 3 to 10 or more. In our case, the data has 6 observations for each of the three variables, which is well within this range.

The panel data regression model is

$$y_{i,t+1} = \alpha_i + X_{i,t}\beta + u_{i,t}$$

for i = 1, …, N; t = 1, …, T. Here, the i subscript denotes the cross-section dimension whereas t denotes the time-series dimension.

There are two main approaches to fitting models using panel data: (1) Fixed-effects model; (2) Random-effects model. The key difference between these two models is how the individual effects ($\alpha_i$) are modeled.

The fixed-effects model is designed to study the causes of changes within a person. It controls for all time-invariant differences between the individuals, so the estimated coefficients of the fixed-effects model cannot be biased because of the omitted time-invariant characteristics like culture, gender, etc. The random-effects model, however, includes time-invariant variables. If we have reason to believe that differences across individuals have some influence on our dependent variable, then we should use random-effects model. In this section, we apply both of these two models and then make a comparison to see which one performs better on the data.

### 3.5.2 Data Transformation

As an extension of previous analysis, we try to figure out what is the exact balance of a customer given the data we have. We have mentioned observations of each month are high correlated for variables repayment status and amount of bill statement. However, there is no obvious correlation between all the three variables repayment status, amount of bill statement and amount of previous payment. Then we could use time invariant variables such as amount of given credit limit, gender, education, marital status and age along with time variant variables repayment status and amount of bill statement to predict another time variant variable amount of previous payment.

Then we modify our dataset. Figure 3.15 below is information of a customer with ID 1. The left 6 variables are time invariant so they stay the same from April to September. The right 3 variables change with month.

| ID | LIMIT_BAL | SEX | EDUCATION | MARRIAGE | AGE | Month | PAY | BILL_AMT | PAY_AMT |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 20000 | 2 | 2 | 1 | 24 | Apr | 2 | 3913 | 0 |
| 1 | 20000 | 2 | 2 | 1 | 24 | May | 2 | 3102 | 689 |
| 1 | 20000 | 2 | 2 | 1 | 24 | Jun | -1 | 689 | 0 |
| 1 | 20000 | 2 | 2 | 1 | 24 | Jul | -1 | 0 | 0 |
| 1 | 20000 | 2 | 2 | 1 | 24 | Aug | -1 | 0 | 0 |
| 1 | 20000 | 2 | 2 | 1 | 24 | Sep | -1 | 0 | 0 |

Figure 3.15

Note that PAY_AMT is amount of previous payment, which is an indicator of previous month. Then we move each PAY_AMT one month up as shown in Figure 3.16. That's the only difference between the two tables shown here. The question mark is what we need to predict: the amount of payment in September.

| ID | LIMIT_BAL | SEX | EDUCATION | MARRIAGE | AGE | Month | PAY | BILL_AMT | PAY_AMT |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 20000 | 2 | 2 | 1 | 24 | Apr | 2 | 3913 | 689 |
| 1 | 20000 | 2 | 2 | 1 | 24 | May | 2 | 3102 | 0 |
| 1 | 20000 | 2 | 2 | 1 | 24 | Jun | -1 | 689 | 0 |
| 1 | 20000 | 2 | 2 | 1 | 24 | Jul | -1 | 0 | 0 |
| 1 | 20000 | 2 | 2 | 1 | 24 | Aug | -1 | 0 | 0 |
| 1 | 20000 | 2 | 2 | 1 | 24 | Sep | -1 | 0 | ? |

Figure 3.16

Our dataset is clearly a panel data after the transformation. We perform linear regression and two panel data analysis models: fixed-effects model and random-effects model on our dataset.

### 3.5.3 Linear Regression

By overlooking all time effects and treat each variable as an independent variable, we could build a linear regression model. The significant variables which pass t-test are payment status, amount of bill statement, amount of given credit limit and education level. The regression function is:

$$PAYAMT = 0.0795 - 0.938 \times PAY + 0.0703 \times BILLAMT + 0.0115 \times LIMITBAL - 0.100 \times FEMALE - 0.736 \times University - 0.426 \times Highschool + 0.342 \times OtherEducation - 0.0747 \times Single + 0.244 \times OtherMarital - 0.00650 \times AGE$$

Before regression, we divide each amount by 1000. It seems that the more months a customer delay, the less he or she will pay next month. As amount of bill statement and credit limit increase, a customer typically pays more next month. Single and female groups tend to pay less next month. The coefficients of education and age are wired.

### 3.5.4 Fixed-Effects Model

Then we build a fixed-effects model. Recall the panel data regression model:

$$y_{i,t+1} = \alpha_i + X_{i,t}\beta + u_{i,t}$$

Note both fixed-effects model and random-effects model are not the same with linear regression since there exist subscript t. The subscript of y is t+1 because we move y to the previous month. In fixed-effects model, there are 3 main characteristics. 1. The intercept in the regression model is allowed to differ among individuals to reflect the unique feature of individual units. We will estimate an $\alpha$ for each y. Since we have total of 13272 customers, there will be 13272 $\alpha$'s to estimate. We could see $\alpha$ as a combination information of all the time invariant variables such as amount of given credit limit, gender, education, marital status and age. 2. Fixed-effects model is appropriate in situations where the individual specific intercept may be correlated with one or more predictors. In this case, it means time invariant predictors may be correlated to time variant predictors. 3. Fixed-effects models are designed to study the causes of changes within a person.

We then figure out the coefficient of payment status and amount of bill statement. They are -1.2999 and 0.2926. Both of them are significant and pass the t-test. Like ordinary linear regression, the signs of coefficients are reasonable. These two coefficients indicate how much y changes overtime, on average per customer, when payment status or amount of bill statement increases by one unit. Also, we estimate an intercept for each individual (13272 in total). Therefore, each individual has its own regression function. Compared with linear regression, it enlarges the effects of these two time variant variables. The absolute values of the coefficients increase.

To compare linear regression and fixed-effect model, we do pFtest. The p-value of this test is $2.2 \times 10^{-16}$, which indicates that fixed-effects model is significantly better than linear regression. Time is critical in our model, and we couldn't treat every time variant variables as independent variables. Panel data analysis is appropriate here. Besides, different backgrounds of customers (shown as the intercepts) play an important role in prediction.

### 3.5.5 Random-Effects Model

Another panel data analysis tool is random-effects model. Unlike fixed-effects model,

$$y_{i,t+1} = \alpha_i + X_{i,t}\beta + u_{i,t}$$

We don't see $\alpha_i$ as parameters to estimate. Instead, we assume that the intercept value of an individual unit is a random drawing from a much larger population with a constant mean. Another two features of random-effects model are listed below. 1. Random-Effects Model is appropriate in situations where the (random) intercept of each cross-sectional unit is uncorrelated with the predictors. In this case, we think "individual" is a random effect. 2. Time-invariant predictors can be used in Random-Effects Model.

The estimated random-effects model is:

$$\begin{aligned} PAYAMT = {} & -0.119 - 1.03 \times PAY + 0.0864 \times BILLAMT + 0.00849 \times LIMITBAL - 0.0303 \\ & \times FEMALE - 0.946 \times University - 0.595 \times Highschool + 0.0369 \\ & \times OtherEducation - 0.107 \times Single + 0.238 \times OtherMarital - 0.00530 \\ & \times AGE \end{aligned}$$

Except for the intercept, the signs of coefficients don't change compared with linear model. We could explain the model in a similar way. However, the coefficients of education and age are still not reasonable.

To decide between fixed or random effects, we run a *Hausman test* where the null hypothesis is that the preferred model is random-effects and the alternative is the fixed-effects. It basically tests whether the unique errors are correlated with the regressors, and the null hypothesis is they are not. The p-value of this test is $2.2 \times 10^{-16}$, which indicates that fixed-effects model is significantly better than random-effects model.

### 3.5.6 Result and discussion

In summary, fixed-effects model performs best among those three models. One of the reasons is maybe error term and the predictors are correlated. In this case, random-effects model estimators will be biased. Individual effects are important and we couldn't treat them as random effects. Since fixed-effects model is better than linear regression, time is also an important factor. Recall the result of fixed-effects model, the coefficient of payment status and amount of bill statement are -1.2999 and 0.2926. Amount of payment will decrease 1299.9 overtime on average per customer, when payment status increases by one unit, and amount of payment will increase 292.6 overtime on average per customer, when amount of bill statement increases by one unit. Other time invariant factors such as amount of given credit limit, gender, education, marital status and age are measured by intercepts, and we could estimate them through the fixed-effects model for each individual.

## 4. Conclusion

We start with logistic regression and train the model on the training dataset: 70% of the raw dataset randomly selected and test the model on the remaining 30%. The logistic regression has achieved 66.98% of accuracy which is better than that of the naïve classifier (50%) where every customer is predicted to default regardless of his background. We plot confusion matrix for the logistic regression and we find the logistic regression tends to wrongly classify not defaulting as defaulting 40% of the time if we predict default. In other words, it has a very high false positive rate. We then look at the default probability distribution under the logistic model, and we find that the logistic model tend to make mistakes around the decision boundaries where points have default probability close to 0.5. There was a high portion of instances in our dataset that seems to cluster around probability 0.5 under the logistic model.

This problem motivates the support vector machine model, which considers distances to the closest points. It is a max margin classifier that are not prone to the influence of outliers but misclassification. We then derive the optimization problem for support vector machine and apply cvx.opt packages in python to solve the quadratic programming in the dual problem. With linear kernel, our support vector machine model achieves an accuracy of 76.49%. Besides overall accuracy, we also show how support vector machine has reduced the false positive rate by 10% under the logistic regression as we have hypothesized early when prompting the support vector machine model.

We then look at another type of linear classifier, regression trees, which is simply a stepwise linear function. We apply pruning, bagging and bootstrapping techniques and finally settle down with a random forest model. We find that the random forest model has improved both true positive and false positive rate. The plot on variable importance is in accordance with the coefficients in the logistic regression, suggesting that previous payment, repayment status and amount of bill statement are the most significant variables to predicting default payment next month. This motivates our panel data model which studies the relationship among these variables. In particular, we found that fixed effect model seems to perform the best among three panel data models.

The random forest model is the best classifier for this dataset, which has an accuracy of 85.08%