Data Science for Business - Team Project Final Report

Hong Anh Chu (hc392), Jiakai Liu (jl1256), Eiad Mohamed (em422), Anshuman Nemali (an297), Caiqing Xie (cx88). Section C - Team 55

Predicting Customer Churn in Credit Card Industry

Target Variable: The primary target variable is 'Exited,' indicating whether a customer has churned (Yes/No). We want to test how our model compares to actual data on whether or not commercial banking customers churn or not.

Data Source: Predicting Churn for Bank Customers

Business Understanding:

Consumer Banking - Credit Card Churn Rates in Spain, France, and Germany

Problem Identification and Motivation:

In today's highly competitive commercial banking landscape, retaining existing customers is often more cost-effective than acquiring new ones. For instance, upfront costs that the bank may spend on new customers include marketing, onboarding, and initial support for new customers. Furthermore, a higher churn rate may not only lead to a decrease in profitability, but may also be symptomatic of deeper issues within the bank's service offerings or customer engagement strategies. This could lead to reputational risks and a potential decline in market share over time.

However, not all churn is equal. Some customers may leave due to reasons beyond a bank's control, such as relocation to an area where the bank does not operate, or death. Others might leave due to solvable issues like dissatisfaction with customer service or lack of personalized offers. Identifying the latter group is essential, as targeted interventions could prevent churn or decrease churn rates in this group of customers. One of the primary concerns for banks, especially in the consumer credit card segment, is the churn rate – the percentage of credit card customers who close their accounts within a specific time frame. Thus, it is crucial to accurately predict which customers are most likely to churn due to preventable reasons, and what reasons they have for churning away to other product offerings.

Data Manipulation & Cleaning:

Data Mining Problem: While the dataset appears relatively clean, there may be challenges in terms of feature selection, handling imbalanced data, and understanding potential multicollinearity among independent variables.

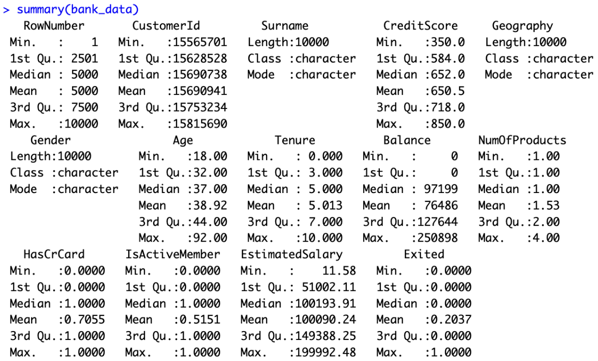
Data Mining Solution: A predictive data mining solution will be implemented to predict the likelihood of a customer churning. By identifying these high-risk customers early on, the bank can take proactive measures, such as offering special incentives or personalized services, to retain them.

Data Understanding:

We believe the dataset is sourced from the bank's internal CRM system. The main focus is on customer profiles and banking behavior with respect to credit card usage.

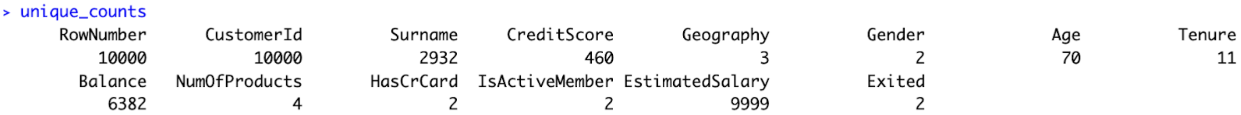
Understanding the raw data set:

1) **Summary** of Data (pre-cleaning):



```
> summary(bank_data)
   RowNumber       CustomerId         Surname           CreditScore     Geography        
 Min.   :    1   Min.   :15565701   Length:10000       Min.   :350.0   Length:10000      
 1st Qu.: 2501   1st Qu.:15628528   Class :character   1st Qu.:584.0   Class :character  
 Median : 5000   Median :15690738   Mode  :character   Median :652.0   Mode  :character  
 Mean   : 5000   Mean   :15690941                      Mean   :650.5                     
 3rd Qu.: 7500   3rd Qu.:15753234                      3rd Qu.:718.0                     
 Max.   :10000   Max.   :15815690                      Max.   :850.0                     
    Gender               Age            Tenure          Balance        NumOfProducts 
 Length:10000       Min.   :18.00   Min.   : 0.000   Min.   :     0   Min.   :1.00  
 Class :character   1st Qu.:32.00   1st Qu.: 3.000   1st Qu.:     0   1st Qu.:1.00  
 Mode  :character   Median :37.00   Median : 5.000   Median : 97199   Median :1.00  
                    Mean   :38.92   Mean   : 5.013   Mean   : 76486   Mean   :1.53  
                    3rd Qu.:44.00   3rd Qu.: 7.000   3rd Qu.:127644   3rd Qu.:2.00  
                    Max.   :92.00   Max.   :10.000   Max.   :250898   Max.   :4.00  
   HasCrCard      IsActiveMember   EstimatedSalary         Exited      
 Min.   :0.0000   Min.   :0.0000   Min.   :    11.58   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 51002.11   1st Qu.:0.0000  
 Median :1.0000   Median :1.0000   Median :100193.91   Median :0.0000  
 Mean   :0.7055   Mean   :0.5151   Mean   :100090.24   Mean   :0.2037  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:149388.25   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :199992.48   Max.   :1.0000  
```

2) Check the **number of unique values** in each column in all columns to see which columns have duplicate unique values (none):

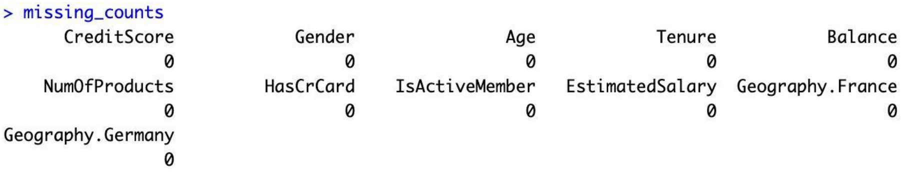


```
> unique_counts
      RowNumber      CustomerId         Surname     CreditScore       Geography          Gender             Age          Tenure
          10000           10000            2932             460               3               2              70              11
        Balance   NumOfProducts       HasCrCard  IsActiveMember EstimatedSalary          Exited
           6382               4               2               2            9999               2
```

3) **Our Target Value**: Y = Exited. Y = 1 means that customers will/have churned, Y= 0 means that customers will not churn.

Understanding the features in the dataset:

1) Check for missing values to make sure there are no null observations.



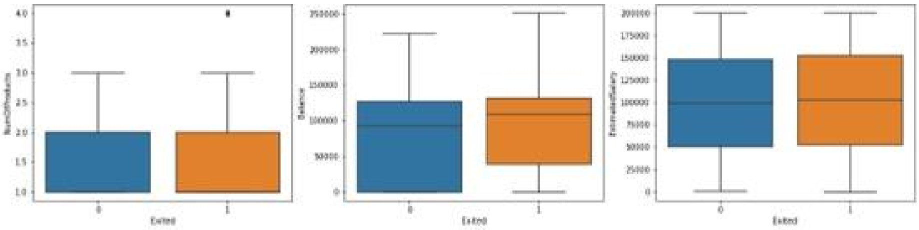
```
> missing_counts
      CreditScore            Gender               Age            Tenure           Balance 
                0                 0                 0                 0                 0 
    NumOfProducts         HasCrCard    IsActiveMember   EstimatedSalary  Geography.France 
                0                 0                 0                 0                 0 
Geography.Germany 
                0 
```

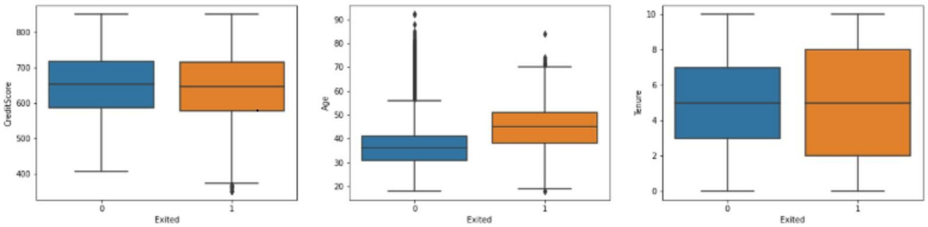
Variable Insights:

- *Continuous Variables*: CreditScore, Age, Tenure, Balance,
  NumOfProducts, and EstimatedSalary.
- *Categorical Variables*: Geography and Gender.
- *Binary Variables:* HasCrCard, IsActiveMember, and Exited.

Understanding numerical features:

Numerical (Continuous) Features: Box plots were utilized for CreditScore, Age, Tenure, NumOfProducts, Balance, and EstimatedSalary to visually inspect their spread and the incidence of any potential outliers.





Box Plots:

<u>CreditScore vs. Exited:</u> The median credit score for both those who exited and those who did not is approximately the same. There are some outliers, particularly for those who exited.

<u>Age vs. Exited:</u> Exited customers tend to be older with a median age higher than those who stayed. The greater spread shows a wider age range for exited customers.

<u>Tenure vs. Exited:</u> Tenure (length of the relationship with bank) is approximately the same for both groups.

<u>NumOfProducts vs. Exited:</u> The number of products used by customers is approximately the same for both groups. There's an outlier for those who exited with a higher number of products.

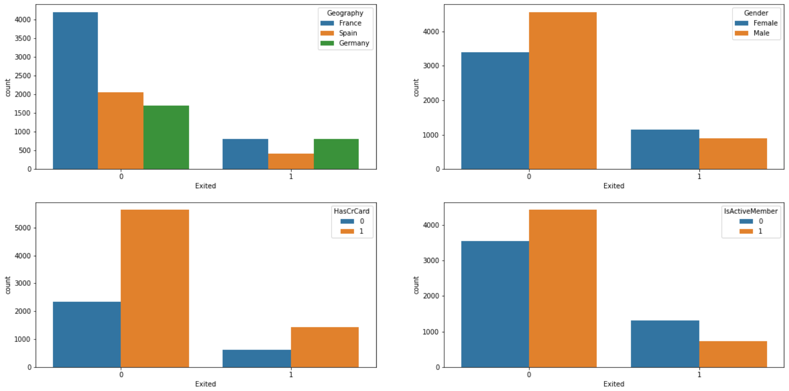
<u>Balance vs. Exited:</u> Customers who exited tend to have a higher median balance compared to those who stayed. There's a significant number of customers who did not exit with a balance of zero.

<u>EstimatedSalary vs. Exited:</u> The estimated salary does not seem to significantly differentiate between those who exited and those who didn't. The medians are closely aligned. From this analysis, variables like 'Geography', 'Gender', 'IsActiveMember', and 'Age' seem to have a more evident relationship with the 'Exited' status.

Understanding the categorical features:

Counts for Countries, Gender, ‘Card Member’, and ‘Has a Credit Card’ for those who Exited (Churned)

Categorical Data: Bar charts were plotted for Geography, Gender, HasCrCard, and IsActiveMember to observe their distributions against the target variable.



Bar Charts:

Geography vs. Exited: The majority of bank customers from France did not churn. The proportion of customers who churned is relatively higher for Germany compared to Spain and France. The number of customers from Spain who churned is lower than those who stayed.

HasCrCard vs. Exited: A significant number of customers who did not churn have credit cards. The difference between customers who churned with or without a credit card is smaller.

Gender vs. Exited: More females churned than males. A higher number of males did not churn compared to females.

IsActiveMember vs. Exited: Most of the active members did not churn. A significant proportion of non-active members churned.

Data Preparation

Data Split

- **75%** = Training Data; **25%** = Testing Data

We are using the package ‘Caret’ to split our data. The random seed is 1256. By specifying a random seed, we can ensure that you obtain the same random results every time you run the same code or algorithm.

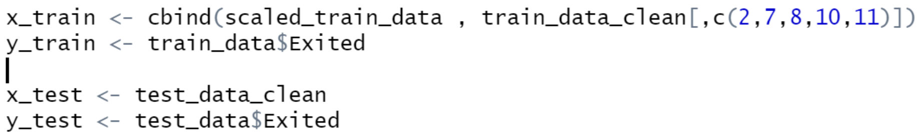
The dataset was divided into training and testing sets using a 75-25% split. This ensures we have a separate dataset to validate our model's performance. After splitting our data set, we have 7501obs of training data and 2499obs of testing data.

Dropping Features: Columns like RowNumber, CustomerId, Surname are the unique identifiers of an individual customer. These features have many unique values so we removed them. Exited is our variable that we are targeting as Y so it is unnecessary to train or test our data on it and thus, it was removed.

One-Hot Encoding: For categorical variables like Geography, one-hot encoding was applied. This process transforms categorical data into a binary matrix, which is easier for models to interpret.

Data Standardization: We found that there is a huge gap between the values in the column Balance. Standardization ensures that all features (variables) are on the same scale. This is important because some machine learning algorithms are sensitive to the scale of the input data. Features on different scales can lead to poor model performance. Further, since we will use the K-Nearest Neighbors Model in our analysis, standardization is particularly important when using distance-based algorithms.

Check and Identify:



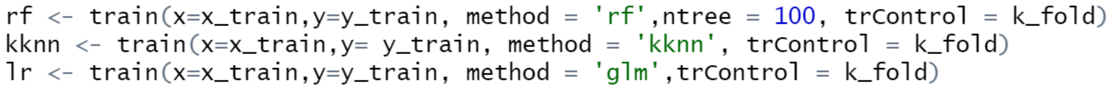
```
x_train <- cbind(scaled_train_data , train_data_clean[,c(2,7,8,10,11)])
y_train <- train_data$Exited

x_test <- test_data_clean
y_test <- test_data$Exited
```

Modeling

K-Fold: We use K-Fold (K=5) cross-validation to help in detecting and mitigating the overfitting of our model on our data. It allows us to assess how well a model applies to different subsets of the data, and by averaging the results, it provides a more accurate representation of a model's performance across the dataset.

Model Building: Using the Caret package, we built 3 models: Random Forest, K Nearest Neighbors and Logistic Regression:

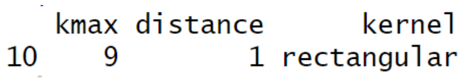


```
rf <- train(x=x_train,y=y_train, method = 'rf',ntree = 100, trControl = k_fold)
kknn <- train(x=x_train,y= y_train, method = 'kknn', trControl = k_fold)
lr <- train(x=x_train,y=y_train, method = 'glm',trControl = k_fold)
```

Model Training and Prediction: We got the highest predicted accuracy from our K Nearest Neighbours with **0.7214886**. Compared to the performance with our training data, the random forest model and logistic regression model had great accuracy with our training data, but performed poorly in predicting our test data. This means that our random forest model and logistic regression model were overfitting our data significantly. In comparison, although the accuracy of K Nearest Neighbours is 0.72, which happens to be overfitting slightly, it performed much better than the logistic regression and random forest models. As a result, we decided to choose K Nearest Neighbors to be our final model.

K Nearest Neighbors: The performance of our KNN model on the training data is below. The best accuracy we were able to achieve was **0.83**.The predictive performance of our model on our testing data is below. The best accuracy we were able to achieve was **0.72**.

Improving our model by finding optimal hyperparameters: Although the accuracy of 0.72 for our model's prediction of the testing data is decent, we still want to try our best to improve our K Nearest Neighbor model. The idea is to find the possible hyperparameters for our K Nearest Neighbor model. The parameters for K Nearest Neighbor are 'kmax', 'distance' and 'kernel'. We built a grid of the possible hyperparameters and used the grid to train our K Nearest Neighbor model to improve it. We find that when 'kmax' equals 9, 'distance' equals 1 and the 'kernel' is rectangular, we can improve our K Nearest Neighbor model.

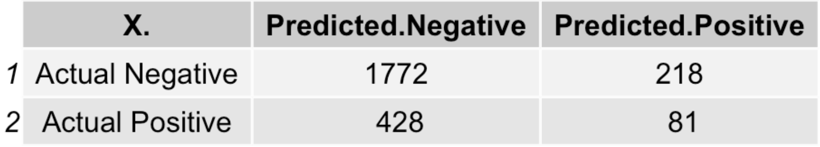


```
   kmax distance      kernel
10    9        1 rectangular
```

After using our improved K Nearest Neighbor model to form a prediction based on our test data, the best accuracy we got is **0.7414866**. This is 2% higher than our previous K Nearest Neighbor model.

Evaluation

Confusion Matrix:



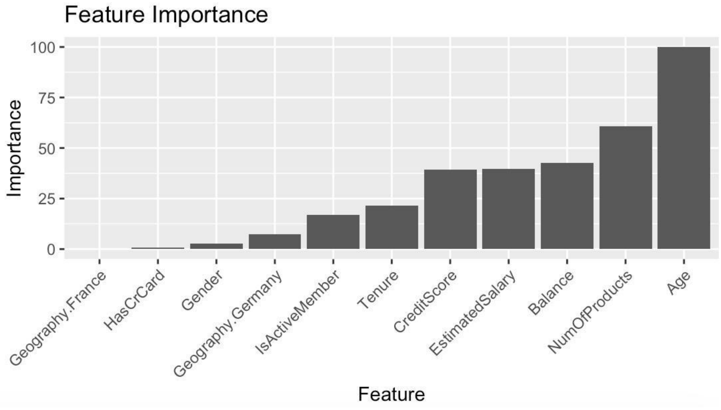
	X.	Predicted.Negative	Predicted.Positive
1	Actual Negative	1772	218
2	Actual Positive	428	81

This matrix provides a summary of the prediction results for any given classification problem. For our model, we found that the number of true negatives is 1772, meaning 1772 instances were correctly predicted as negative. The number of false positives is 218, indicating 218 instances were incorrectly predicted as positive when they were actually negative. The number of false negatives is 428, which means 428 instances were incorrectly predicted as negative when they were actually positive.

The number of true positives is 81, so 81 instances were correctly predicted as positive. Based on this, we can compute several diagnostic metrics like accuracy, precision, and recall to understand the model's performance better.

Deployment

Feature Importance: Although an improved K Nearest Neighbors is the best model for us, it is hard to find feature importance directly through KNN, we used the random forest model to identify the importance of different features in the data set.



This chart displays the significance of various features used in a predictive model. The height of the bars represents the relative importance of each feature. As we can see, the most important feature, according to the chart, is Age, followed by NumOfProducts and Balance. Features like Geography: France and HasCard have the least importance in this model.

Deployment in Business:

The results of our data mining endeavor can be used to inform banks as to how to reduce or eliminate churn rates amongst its credit card customers in three key geographies: France, Spain, and Germany. By using our KNN model, we can say that our model is roughly 74% accurate in predicting churn rates among customers. Through our Primary Component Analysis (PCA) in our Random Forest model, we suggest that commercial banks should make it a priority to target older customers, increase the number of credit products available, reduce outstanding balances, and seek individuals with higher estimated salaries.

One issue that becomes salient through our analysis is that controlling for customers' balances is very difficult to do, especially given the fact that a bank cannot simply reduce/eliminate a customer's credit card debt. As such, focusing on other important features is crucial. Another issue is that our study may not translate to non-EU countries where commercial banking systems vary. From an ethical standpoint, we have eliminated privacy concerns in our analysis. Given the importance of data privacy in Europe after the passing of General Data Protection Regulation (GDPR) legislation, we made sure to remove any personal identifier data from our dataset, such as 'CustomerId' and 'Surname'.

Overall, our model does a decent job of predicting churn rates among credit card customers in Germany, Spain, and France, however, a significant margin of error remains. We would urge our client, the commercial banks, to be wary of this error, and to continue to innovate to create commercial credit products with robust rewards or points programmes and to invest heavily in understanding their consumer base better. Since age seems to be the biggest indicator of churn, creating coupon plans, loyalty rewards, or deals through third-party vendors could be other avenues to pursue to prevent the elderly from churning.

Appendix I - Qualitative Analyses:

Spain: Traditionally, Spain's banking sector has been characterized by its large number of branches per capita. However, with the advent of digital banking, the emphasis has shifted to online and mobile services, with ATMs and bank branches being closed around the country.[1] While credit cards are commonly used, Spain is still a country that relies on cash for financial transactions.[2]

France: French consumers are known for their cautious approach to credit. The majority prefer using debit cards over credit cards, perhaps because it allows greater control over expenses and the minimizing of debt.[3] Still, the credit card market remains substantial, making customer acquisition and retention strategies vital for banks.

Germany: The German banking system is one of the largest in the world and is known for its diversity and competition. Historically, Germans have shown a preference for cash transactions, but there's a steady rise in credit card usage, especially among younger demographics and frequent travelers.[4]

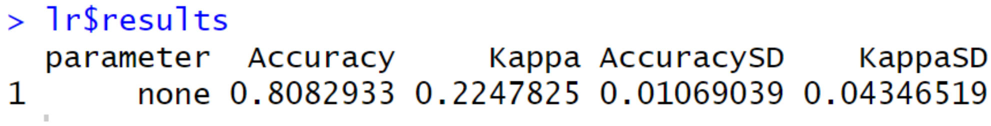
For banks in these regions, understanding churn is imperative. A high churn rate could indicate customer dissatisfaction, better offers from competitors, or a mismatch between the bank's services and customer needs. Moreover, acquiring a new customer is often more expensive than retaining an existing one, making churn reduction a top priority.

[1] Traditional Commercial Banking - Spain | Market Forecast
[2] Spain Cards and Payments - Opportunities and Risks to 2026
[3] Country survey France: cash displacement remains a work in progress
[4] A perspective on German payments

Appendix II - Linear Regression:

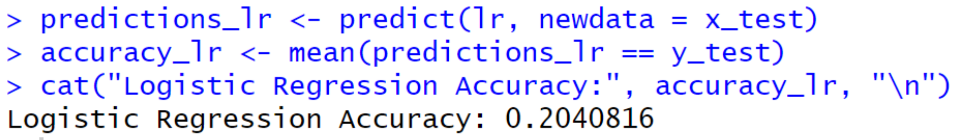
Logistic Regression: The performance of our model on the training data is as below.

The best accuracy we were able to achieve was **0.83.**



```
> lr$results
  parameter  Accuracy     Kappa AccuracySD    KappaSD
1      none 0.8082933 0.2247825 0.01069039 0.04346519
```

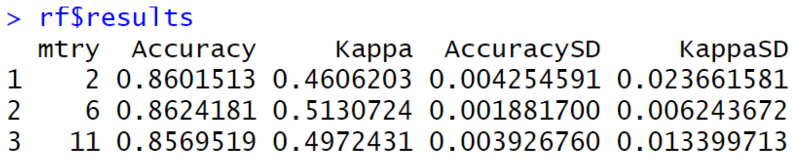
The predictive performance of our model on our testing data is below. The best accuracy we were able to achieve was **0.20**.



```
> predictions_lr <- predict(lr, newdata = x_test)
> accuracy_lr <- mean(predictions_lr == y_test)
> cat("Logistic Regression Accuracy:", accuracy_lr, "\n")
Logistic Regression Accuracy: 0.2040816
```

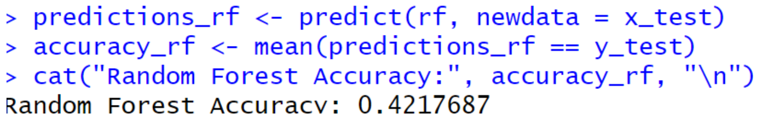
Appendix III - Random Forest Model:

Random Forest: The performance for training data are as below. We got the best accuracy for 0.86



```
> rf$results
  mtry  Accuracy     Kappa  AccuracySD     KappaSD
1    2 0.8601513 0.4606203 0.004254591 0.023661581
2    6 0.8624181 0.5130724 0.001881700 0.006243672
3   11 0.8569519 0.4972431 0.003926760 0.013399713
```

The prediction of test data is as below. We got the accuracy for 0.42.



```
> predictions_rf <- predict(rf, newdata = x_test)
> accuracy_rf <- mean(predictions_rf == y_test)
> cat("Random Forest Accuracy:", accuracy_rf, "\n")
Random Forest Accuracv: 0.4217687
```

Appendix IV - Confusion Matrix Inferences:

Accuracy, Precision, and Recall:

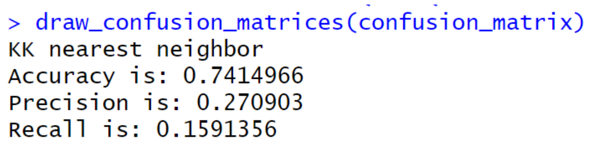
This helps us understand the Confusion Matrix and what it tells us about our model's performance on the data.

TruePositive: correctly identified churn rates, Precision (PPV, positive predictive value): TruePositive / (TruePositive + FalsePositive)

Total number of true positive churners divided by the total number of predicted churners. High Precision means few false-positives: not many return customers were predicted to be customers who churn.

Recall (Sensitivity, Hit Rate, True Positive Rate): TruePositive / (TruePositive + FalseNegative)

Predict the most positive or churning customer correctly. High recall means low false-negative, which means few churning customers were predicted as return customers.



```
> draw_confusion_matrices(confusion_matrix)
KK nearest neighbor
Accuracy is: 0.7414966
Precision is: 0.270903
Recall is: 0.1591356
```

For our K Nearest Neighbor model, the accuracy represents the proportion of correctly predicted samples to the total number of samples. In our case, our model's accuracy is 0.7414966, which means the model correctly predicted approximately

74.15% of the samples.

Precision represents the proportion of samples predicted as the positive class (Churn) that are actually part of the positive class. In your example, Precision is 0.270903, indicating that about 27.09% of the samples predicted to churn are truly churning.

Recall represents the proportion of actual positive class samples that the model successfully predicted as the positive class. In your example, Recall is 0.1591356, indicating that the model successfully captured about 15.91% of the positive class samples.

Appendix V - Group Contributions:

- **Coding**: Jiakai Liu, Caiqing Xie
- **Report**: Anshuman Nemali, Caiqing Xie. Edits: Jiakai Liu, Eiad Mohamed
- **Slide Deck:** Hong Anh Chu, Edits: Anshuman Nemali
- **In-Class Presentation:** Anshuman Nemali, Eiad Mohamed, Jiakai Liu