Predicting Customer Churn in the Credit Card Industry

Section C - Team 55

Hong Anh Chu (hc392), Jiakai Liu (jl1256), Eiad Mohamed (em422), Anshuman Nemali (an297), Caiqing Xie (cx88)

The Team





Hong Anh Chu

Jiakai Liu

Eiad Mohamed

Anshuman Nemali

Caiqing Xie

Executive Summary

- Data Mining is the solution to predicting customer churn rate.
  - Our data mining results offer insights for banks to reduce credit card customer churn rates in France, Germany, and Spain.
- Tailoring strategies that reduce churn risk by enhancing retention and boosting brand loyalty.
- Out of all 3 predictive modeling methods we tried,
  - KNN Neighbor gives the highest accuracy prediction.
  - After improving the model, we had higher accuracy (increase of 2%).
- Addressing customers’ balances is a challenging task due to the inability to reduce/eliminate customers’ credit card debt, highlighting the importance of controlling other features.
- While our model effectively predicts customer churn rates in Germany, Spain, and France, there is still a significant margin of error, emphasizing a need for commercial banks to innovate products with robust rewards programs and invest in understanding their customer base better.

Why Predict Customer Churn Rate?

Business Problems

- What factors play a significant role in determining if a customer will churn?
- Is there any regional variance in churn rates? If so, do the determining factors differ by region?

Potential Outcomes

- These insights will enable banks to make accurate predictions on potential churners
- Leading to tailored strategies that:
  - Reduce churn risk (Increase retention)
  - Increase revenue

Target Variable

- The primary target variable is 'Exited,' indicating whether a customer has churned (Yes/No).

Customer Churn Rate Prediction Saves Costs

Challenge 1

Acquiring new customer costs more than retaining old customers

- Cost-effective
- Competitive edge
- Preventable

Challenge 2

Credit card industry in Spain, France, and Germany

- Largely cash-based transaction systems
- Increased adoption of credit cards
- Rise of digital banking/payments

Challenge 3

High churn rate implications

- Customer dissatisfaction
  - (Mismatch between customer and product or service offering)
- Competitive offers elsewhere

Data Mining Is The Solution to Predicting Churn Rate

Early identification of high-risk customers allows the bank to proactively retain them through incentives and personalized services

- Age, number of banking products, account balance, etc., affect churn rate
- These insights help banks:
  - Tailoring services and offers
  - Customer engagement strategies
  - Foster brand loyalty

Data Mining Road Map

Understanding the raw dataset

1. Data summary
2. Find unique values
3. Get target value Y

Understanding the features

1. Find missing values
2. Interpret insights
   a. Numerical features
   b. Categorical features

Data Preparation

1. 25/75 data split
2. Dropping feature(s)
3. One-hot encoding
4. Data Standardization
5. Checking & Identifying

Modeling

1. K-Fold
2. Model Building
3. Train & Predict
4. RF, KKN neighbors,, log regression
5. Model Improvement
6. Evaluation
7. Confusion matrix
8. Precision, Recall, Accuracy

Deployment

1. Feature Importance
2. Business application
3. Ethical considerations
4. Associated risk(s)

Data Summary Provides Understanding of Data Set

```
> 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  
```

Numerical & Categorical Gives Valuable Insights

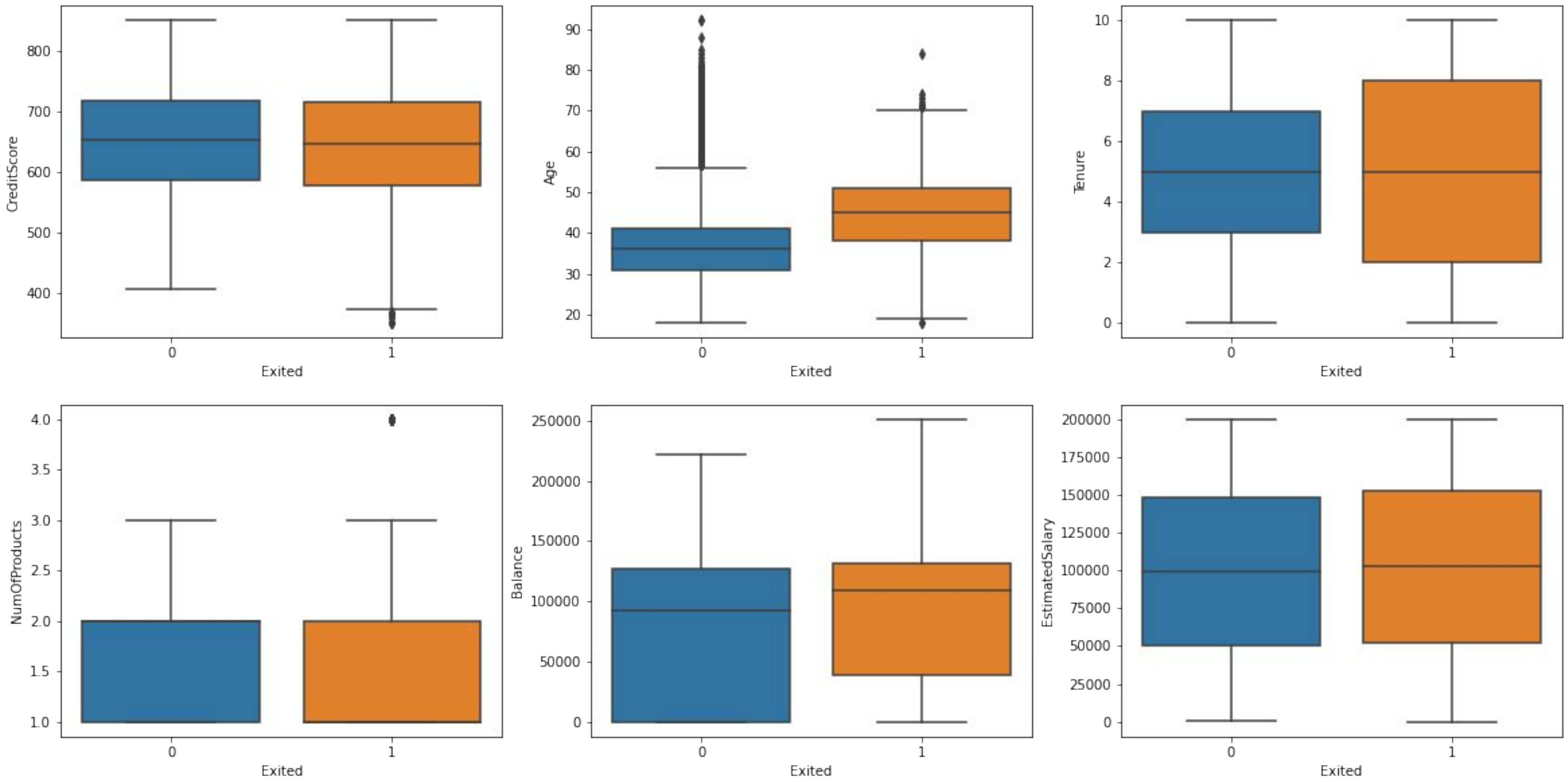
1. Finding Missing Values

```
> missing_counts
      RowNumber      CustomerId         Surname     CreditScore       Geography          Gender 
              0               0               0               0               0               0 
            Age          Tenure         Balance   NumOfProducts       HasCrCard  IsActiveMember 
              0               0               0               0               0               0 
EstimatedSalary          Exited 
              0               0 
```

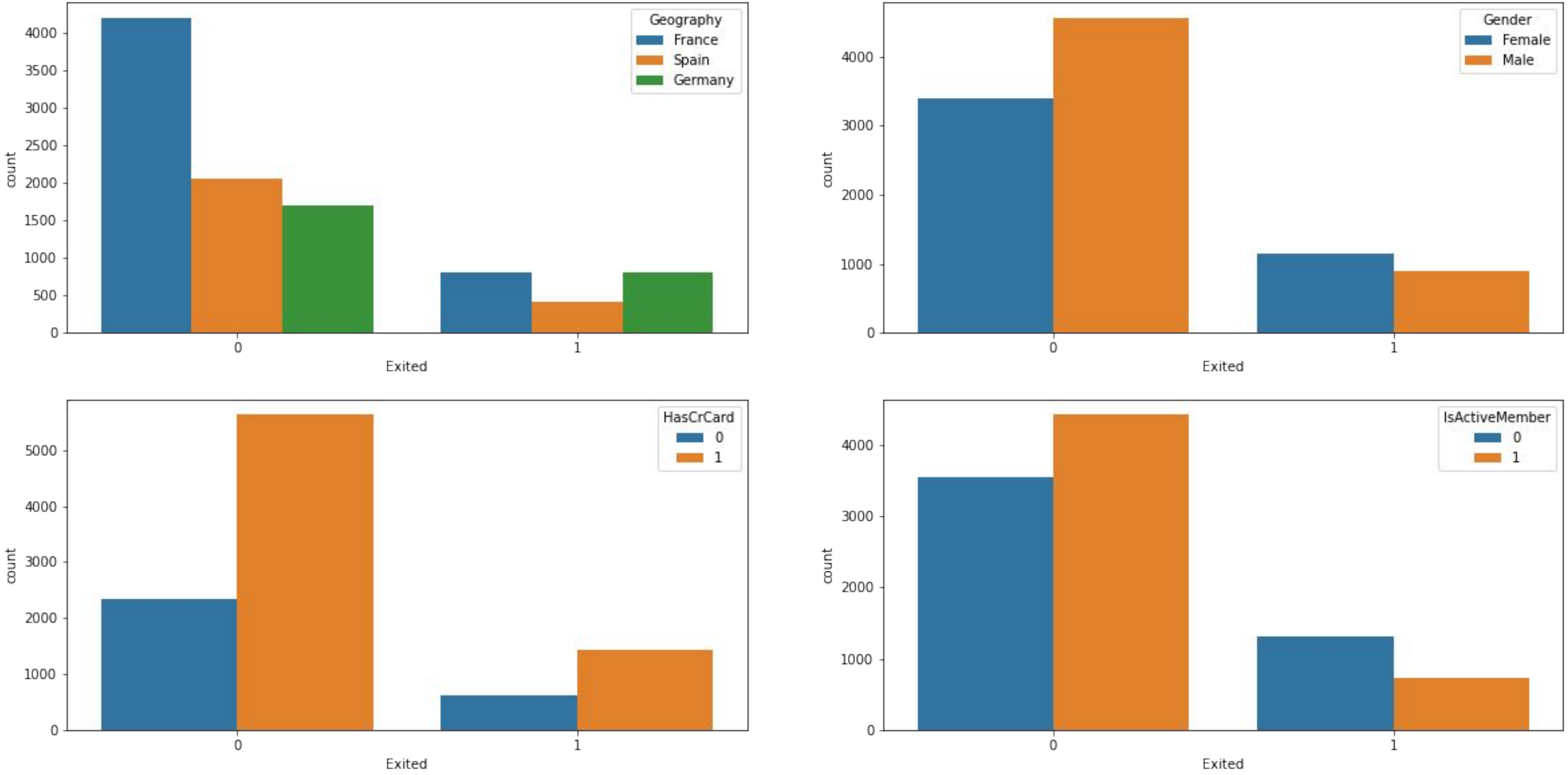
2. Variable Insights

- **Numerical features:**
  - CreditScore, Age, Tenure, NumOfProducts, Balance, EstimatedSalary
- **Categorical features:**
  - Geography, Gender, HasCrCard, IsActiveMember
- A deeper dive into individual features was performed to understand their distribution and relation to the target variable, 'Exited'.

Exited (Churners) Compared on Credit Score, Age, Tenure, Balance, and Estimated Salary



Observing Geography, Gender, HasCrCard, and IsActiveMember's distributions against Exited



KK-Nearest Neighbors Model Gives the Highest Accuracy

- K Nearest Neighbors

```
> cat("K Nearest Neighbors Accuracy:", accuracy_kknn, "\n")
K Nearest Neighbors Accuracy: 0.7214886
```

- Random Forest

```
> cat("Random Forest Accuracy:", accuracy_rf, "\n")
Random Forest Accuracy: 0.4057623
```

- Logistic Regression

```
> cat("Logistic Regression Accuracy:", accuracy_lr, "\n")
Logistic Regression Accuracy: 0.2040816
```

After Improving, KKN Neighbors Gives the Highest Accuracy, 2% Higher Than the Previous Model

Finding **optimal hyperparameters** (kmax, distance, and kernel):

```
library(kknn)
param_grid <- expand.grid(kmax = c(5, 7, 9),distance = c(1,2), kernel = c("optimal", "rectangular"))
kknn_model <- train(
  x = x_train,
  y = y_train,
  method = "kknn",
  tuneGrid = param_grid,
  trControl = k_fold)
print(kknn_model$bestTune)
```

When **kmax = 9**, **distance = 1,** and **kernel = rectangular** → The **best** kernel kk-near neighbor model:

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

The **best accuracy** is **0.7414866** & is **2% higher** than the previous kernel k near neighbor model

```
> predictions_kknn <- predict(best_kknn_model, newdata = x_test)
> accuracy_kknn <- mean(predictions_kknn == y_test)
> cat("Best KK Nearest Neighbors Accuracy:", accuracy_kknn, "\n")
Best KK Nearest Neighbors Accuracy: 0.7414966
```

Confusion Matrix Gives Precision, Recall, & Accuracy

1. **PRECISION = 0.270903**
 - Indicating that about **27.09%** of the samples predicted as Churn are truly Churn.

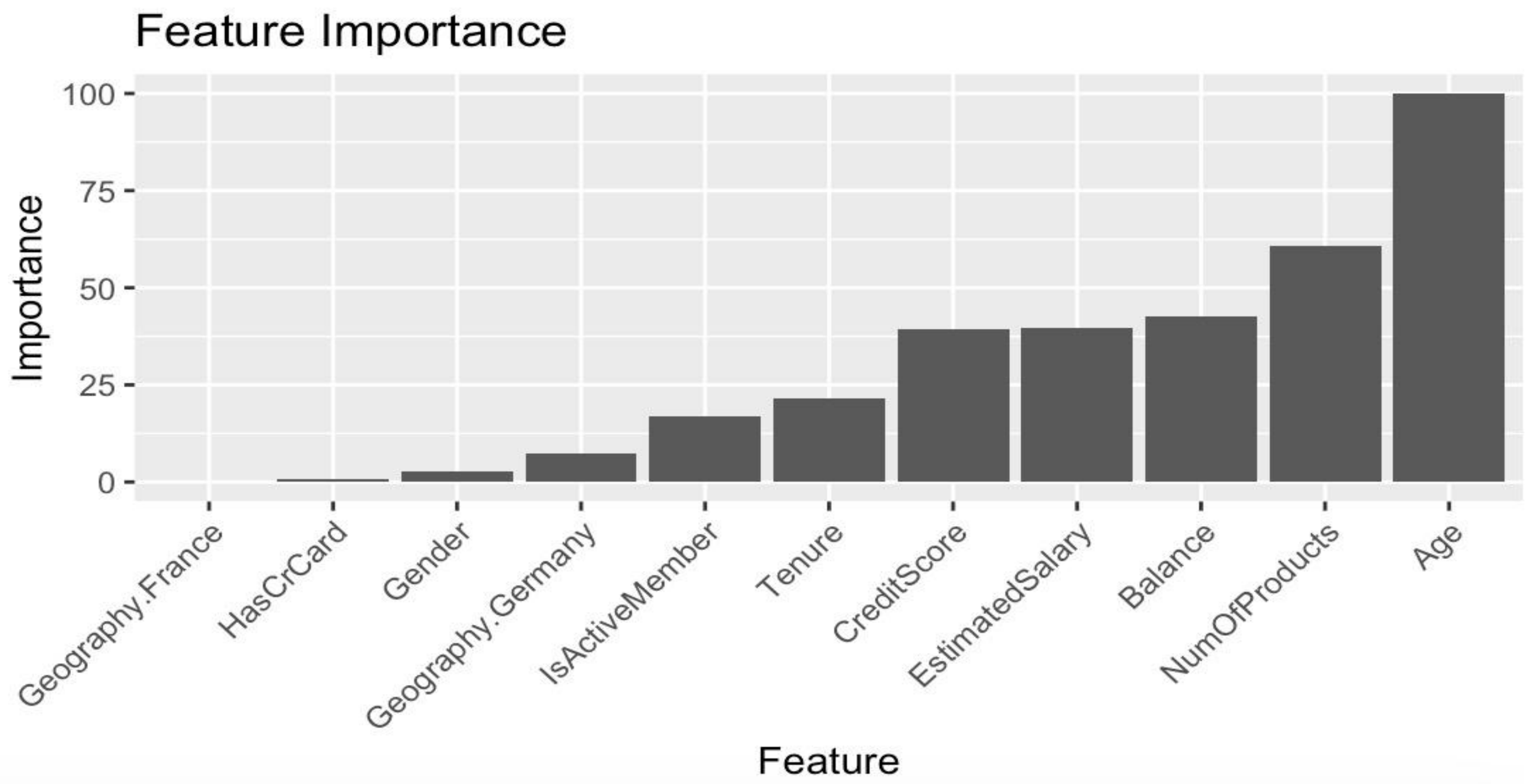
2. **RECALL = 0.1591356**
 - Indicating that the model successfully captured about **15.91%** of the positive class samples.

3. **ACCURACY = 0.7414966**
 - Indicating that the model correctly predicted approximately **74.15%** of the samples.

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

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

The most important features are Age, NumOfProducts, and Balance



Business Deployment, Ethical Concerns, and Risks

<u>Key Takeaway</u>: *Prioritize targeting older customers.*

How? *Increasing the number of credit products available and seeking customers with higher estimated salaries.*

<u>Business Deployment Consideration:</u>

- Addressing customers' balances is a challenging task due to inability to reduce/eliminate customers' credit card debt, highlighting the importance of other features.

<u>Ethics:</u>

- Addressed privacy concerns by removing personal identifiers such as "CustomerID" and "Surname", in compliance with Europe's General Data Protection Regulation (GDPR).

<u>Risks Associated:</u>

- While our model effectively predict churn rates, there's still a significant margin of error, emphasizing a need for banks to innovate products and invest in understanding customer base better.
- Limited applicability to non-EU nations with different commercial banking systems.

Thank you!

Any questions?