Customer Churn Analysis in Banking Sector

PROJECT REPORT

GROUP 8

Alok Singh (50545018)
Aseem Salim (50545729)
Dhiraj Sanjay Landge (50479342)
Rohan Sharma (50545225)
Shriram Sankaranarayanan (50548133)

ABSTRACT-

This research project aims to analyze and address customer churn in the banking industry by identifying the key predictors influencing churn using predictive analytics model. The database contains detailed information about 10,000 bank customers, including their financial health and personal information. The survey examines borrowing costs, demographics, bank balances, infrastructure, and other relevant factors. Key research questions focus on the likelihood of customer convergence, predicting the impact of future trends, proposing strategies to reduce customer retention, and identifying the costs of building customer availability emphasis on preserving existing ones rather than acquiring new ones.

Predictive analytics models including logistic regression, decision trees, random forests, and neural networks will be used to predict customer churn The model includes customer data such as credit score, age, spending, product role, credit card ownership, level of activity, and salary expectations. Significance analysis of these predictive models should be used to determine which predictors have the greatest impact on churn. Machine learning models will rank the features based on their contribution to the predictive power of the model. The ultimate objective of this research is to improve client relationships, engagement tactics, and fundamental performance of banks while reducing acquisition costs.

Keywords- Customer churn, Banking industry, Predictive analytics models, Retention strategies, Churn predictors.

PROBLEM DEFINITION-

Customer churn in banking applies to the occurrence in which a bank has lost customers for a variety of reasons. This can occur for a variety of causes, including unhappiness with the bank's services, better offers or services from rivals, changes in client requirements or financial conditions, or a general degradation in the bank's brand impression. Because recruiting new clients is often more expensive than maintaining current ones, churn may have a major impact on a bank's income and market share. Furthermore, high turnover rates can have a negative influence on a bank's reputation and client loyalty, making it critical for banks to recognize and address the root causes of churn in order to retain a stable and pleased customer base.

PROBLEM STATEMENT-

The banking industry is grappling with a fundamental challenge—customer inflow, where customers terminate their ties with banks for a variety of reasons. This report addresses this issue by probing key research questions aimed at understanding, predicting and reducing customer visits. The following questions guide our research-

1. **Find the chances of whether a customer will churn or not: -**
   How can we know how likely it is that customers will actually come on board? This report focuses on developing predictive models to identify and measure customers at risk of leaving the bank.

2. **Find predictors which are influencing customer churn: -**
   What are the factors that most affect customers' access to the banking industry? We aim to report a range of forecasts, including financial health, demographics and banking behavior, that play an important role in customer churn.
3. **What can be done to avoid customer churn? -**
   What strategies can be implemented to restrict customer flow? This report delves into the root causes of churn and offers effective strategies to increase customer satisfaction and loyalty.
4. **How to retain existing customers as getting new customers have more cost associated with compared to retaining the existing customers: -**
   Considering that acquiring new customers is often costly, how can we strategically retain existing customers? Our research analyses the financial implications of retaining current customers and formulates strategies for developing optimal products for long-term customer relationships.

Through comprehensive research and detailed analysis of customer data, this report aims to empower the banking industry to proactively manage customer relationships, improve retention efforts, and establish customers churn impact on efficiency. The findings presented here provide valuable insights for developing targeted strategies to manage customer inflows, ultimately contributing to the continued success of banks.

DATA COLLECTION-

The dataset used for customer churn analysis in banking sector is from Kaggle. This database, which originates from Kaggle, is considered a curated collection, possibly compiled from multiple banking organizations for research and evaluation purposes. The collection contains a wealth of consumer-focused information, from indicators of financial health such as credit scores to personal data such as age, gender, and location. Individual records are uniquely identified by key identifiers such as CustomerID and RowNumber. The 'Exited' variable serves as the primary evaluation point, indicating whether the customer is churned or not (1 for yes, 0 for no). However, this data is a good source for researching the dynamics of customer churn in the banking industry.

DATA DESCRIPTION-

The 'churn.csv' dataset is a set of data used primarily for to get know to customer will churn or not. It contains detailed information about individual customers, as well as a record of their interactions with financial institutions. Fields such as 'RowNumber' and 'CustomerId' in the dataset mean unique IDs for each customer. The detailed information offered includes 'Title,' 'Accreditation Score,' 'Geography,' 'Gender,' 'Age,' and 'Time of Use,' which are utilized for consumer demographic and financial profiling. To understand customer behavior, financial attributes such as 'Balance', 'NumOfProducts', 'HasCrCard', 'IsActiveMember', and 'EstimatedSalary' are incorporated. These data pieces are critical for institutions to personalize their services and offerings to the specific demands of their customers. The 'Output' column,

which indicates customer turnover, is critical for projecting retention rates, same as organizations estimate service results based on customer history and interaction patterns.

Here's an overview of the dataset:

Column Name	Description
RowNumber	Sequential row numbers
CustomerId	Unique identifier for each customer
Surname	Last name of the customer
CreditScore	Credit score of the customer
Geography	The country of the customer
Gender	Gender of the customer
Age	Age of the customer
Tenure	Number of years the customer has been with the company
Balance	Account balance of the customer
NumOfProducts	Number of products the customer is using
HasCrCard	Indicates whether the customer has a credit card (1) or not (0)
IsActiveMember	Indicates whether the customer is an active member (1) or not (0)
EstimatedSalary	Estimated salary of the customer
Exited	Indicates whether the customer has exited/left (1) or not (0)

The following columns from the dataset are Numerical variables:

Numerical variables
RowNumber
CustomerId
CreditScore
Age
Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
Exited

The following columns from the dataset are Categorical variables:

Categorical variables
Surname
Geography
Gender

The dataset gives us a detailed look at 10,000 bank customers. It tells us about their financial health and personal details. For instance, their credit scores (a measure of financial trustworthiness) range from 350 to 850, with most people scoring around 650.53. This shows us that the bank has customers of all financial backgrounds.

Most of these customers are from France, and there are slightly more men (5,457) than women. The ages of these customers vary a lot too, from young adults at 18 years old to seniors at 92, but most are around 39 years old. Their bank balances are different as well; some don't have money in their accounts, while others have up to $250,898.09. On average, they have about 76,485.89 in their accounts. Most customers use about one or two of the bank's services, showing that the bank might offer them more services.

A good number, 70.55% to be exact, has a credit card from the bank. But only half of the total customers are active, meaning the bank might need to engage them more. Their salaries, or what they might earn, range widely from 11.58 to 199,992.48, with the average being around 100,090.24, showing us that the bank's customers come from all walks of life. A point of concern is that out of all these customers, 20.37% have left the bank, showing the need for the bank to keep its customers better.

We could use predictive analytics models like logistic regression, decision trees, or more advanced machine learning models like random forests or neural networks to determine the likelihood of customer churn. These models would use customer data such as credit score, age, balance, number of products used, credit card ownership, activity level, and expected salary to forecast churn.

The result would be a score that categorizes clients as likely to churn or likely to stay. Identifying the most influential predictors of churn involves feature importance analysis within our predictive models. We can use machine learning models to provide us with a ranking of features based on how much they contribute to the model's predictive power.

DATA OVERVIEW-

```
# Extracting csv data from churn.csv to df0
df0 = read.csv("churn.csv",sep = ",",header = TRUE)
dim(df0) # 10000 rows and 14 variables.
summary(df0)
str(df0) # There are 10000 observations with 14 variables.
```

The presented R code reads data from the CSV file 'churn.csv' into the data frame df0, using a comma as the field separator and considering the first row as headers. It then obtains the dimensions of the data frame, which has 10,000 rows and 14 variables, and displays a summary and structure of the data, providing a brief overview and statistical summary of each column inside the data frame.

DATA PREPROCESSING-

```
# Data Pre-processing
# -------------------

# Check number of unique values in each column.
apply(df0, MARGIN = 2, function(x) length(unique(x)))

# Remove RowNumber and CustomerId as those are unique for each rows
df1 = df0
df1$RowNumber = NULL
df1$CustomerId = NULL

# Remove Surname as it's not relevant
df1$Surname = NULL

# Display number of unique values in each column.
apply(df1, MARGIN = 2, function(x) length(unique(x)))

str(df1) # Now there are 10000 observations with 11 variables.


# Check for missing values

columns <- colnames(df1)
columns
# Convert the blank spaces as NA
for (i in columns) {
  df1[[i]] <- ifelse(trimws(df1[[i]]) == "",
                     NA, df1[[i]])
}

sum(is.na(df1) == TRUE) # Get total number of missing values
colSums(is.na(df1)) # Find missing values in each column
# There are no missing values.
```

A number of processes are conducted during the data pre-processing stage for a banking dataset to assure the data's quality and relevance. The apply() function in R is used to validate the uniqueness of values in each column. This is critical for identifying columns with high variability that may be significant for analysis.

Following the evaluation of the unique values, certain columns are deleted from the dataset to simplify the study. The 'RowNumber' and 'CustomerId' columns are removed since they include unique identifiers for each row and are thus useless for predictive modeling or analysis. Similarly, the 'Surname' item has been eliminated because it is unrelated to the examination of customer attrition in banking.

After these columns are removed, the dataset is reviewed to validate the number of unique values in each remaining column. This stage is critical for understanding the dataset's variability and distribution after first cleaning.

The next crucial step is to check for missing values, which can have a substantial influence on the quality of the study. To effectively identify missing values, blank spaces in R are first transformed to NA (Not Available). This translation facilitates the counting and management of missing data. The overall number of missing values is computed, as well as a column-by-

column count of missing data. In this specific situation, there are no missing values in the dataset.

This rigorous data pre-processing procedure is critical for preparing the dataset for subsequent analysis, such as developing prediction models for customer attrition. It ensures the analysis's dependability and correctness, resulting in more significant insights for banking decision-making.

```
df2 = df1


# Outlier detection using boxplots for continuous variable
num_cols = c("CreditScore", "Age", "Tenure", "Balance", "EstimatedSalary")
cat_cols = c("Geography","Gender","NumOfProducts","HasCrCard","IsActiveMember")
out_col = c("Exited")

op = par()

par(mfrow = c(2,3))

for (i in num_cols) {
  boxplot(df2[i], ylab = i)
}

mtext("Outlier detection for continuous variables",
      side = 3, line = - 2, outer = TRUE)

par(op)

# Outliers present in below columns
# CreditScore, Age

#
par(mfrow = c(1,1))
out <- boxplot.stats(df2$CreditScore)$out
boxplot(df2$CreditScore,
        ylab = "CreditScore"
)
mtext(paste("Outliers: ", paste(out, collapse = ", ")))

# CreditScore Outliers
# 376 376 363 359 350 350 358 351 365 367 350 350 382 373 350
```

```
out2 <- boxplot.stats(df2$Age)$out
boxplot(df2$Age,
        ylab = "Age"
)
mtext(paste("Outliers Range: ", paste(range(out2), collapse = " - ")))

# Age Outliers Range from 63 to 92. 359 Outliers present


# Convert chr variables to qualitative variables
# Geography, Gender
df3 = df2
df3$Geography = factor(df3$Geography)
df3$Gender = factor(df3$Gender)

# Convert num columns having less than 5 distinct values
conv <- sapply(df3, function(x) is.numeric(x) && length(unique(x)) < 5)
df3[conv] <- lapply(df3[conv], as.factor)

str(df3)
```

The focus of the next step of data pre-processing switches to recognizing and dealing with outliers, particularly in continuous variables, and transforming character variables to qualitative variables in a financial dataset.

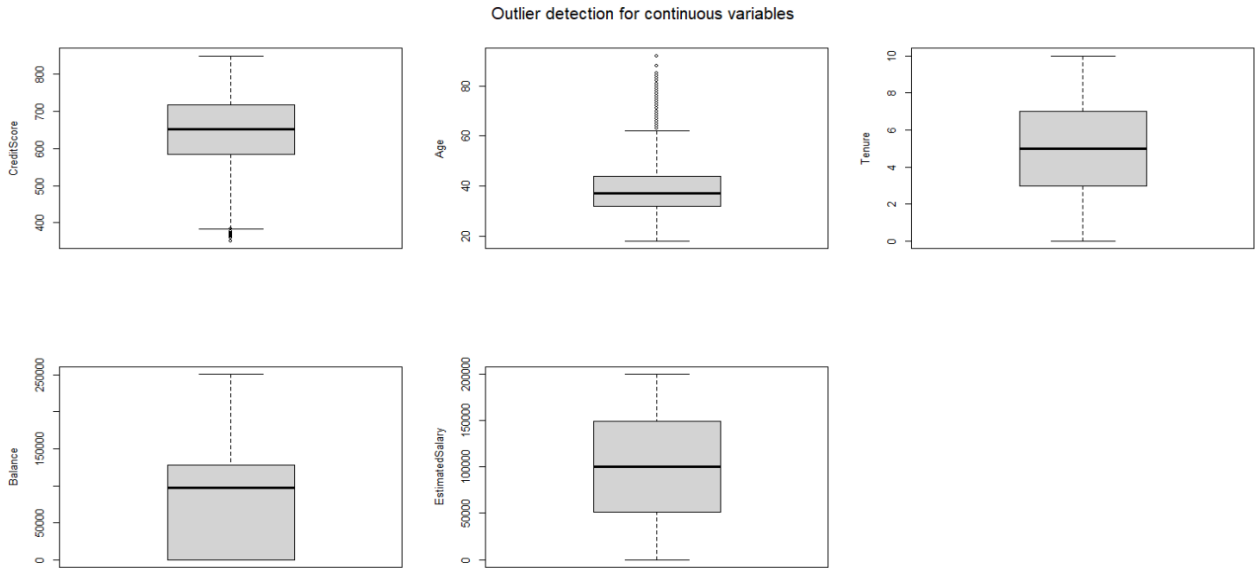
For continuous variables such as 'CreditScore,' 'Age,' 'Tenure,' 'Balance,' and 'EstimatedSalary,' outlier identification is initially conducted using boxplots. Boxplots depict the distribution of these data visually, showing any potential outliers. Outliers can have a substantial impact on the outcomes of any statistical research or predictive modeling.

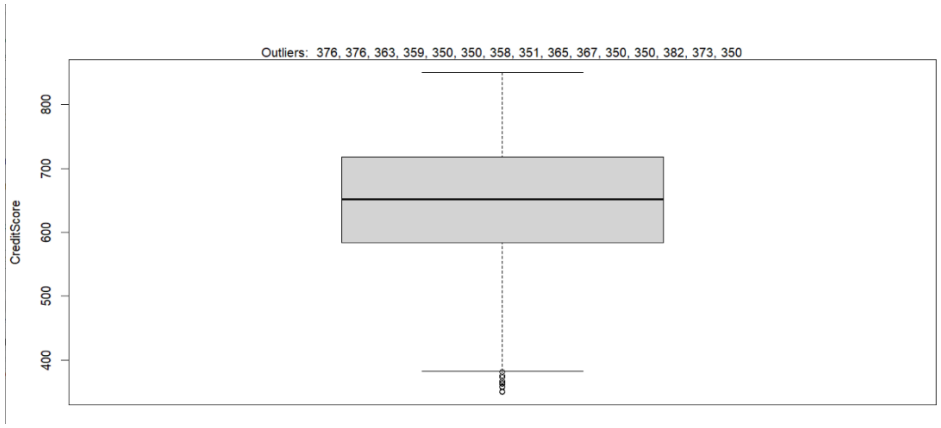
Outliers in 'CreditScore' and 'Age' are shown by the study. Individual boxplots for these two variables are produced to better explore these outliers. The exact outlier values for 'CreditScore' are detected and shown. These are numbers that differ greatly from the remainder of the data points, such as 376, 376, 363, and so on. Similarly, the range of outlier values for 'Age' is determined, ranging from 63 to 92, with a total of 359 outliers.

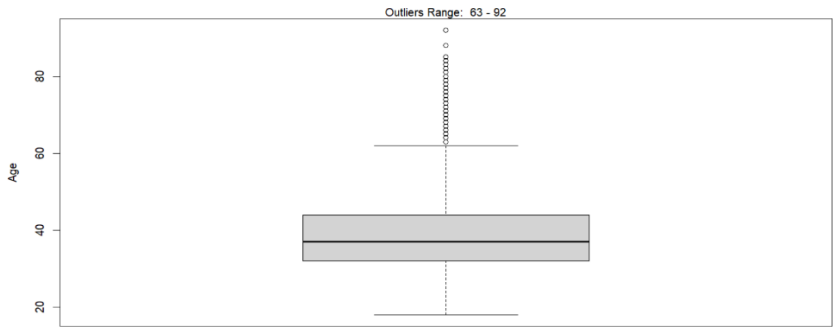
The procedure then continues on to translating character data to qualitative variables after the outlier analysis. 'Geography' and 'Gender' are character variables in this dataset that have been turned into factors (qualitative variables). This conversion is necessary for statistical modeling since it allows these variables to be used in categorical data analysis.

Numeric columns that have less than five different values are likewise transformed to factors. This is done because such numerical variables frequently reflect category data and are best examined as factors. This conversion guarantees that these variables be utilized correctly in later studies, such as logistic regression or decision trees, which deal with categorical data differently than continuous variables.

Following these adjustments, the dataset's structure is reviewed to confirm the changes. The dataset is now made up of a combination of continuous and categorical variables that have been properly organized for sophisticated data analysis operations. This organized approach to data pre-processing is critical for assuring the correctness and efficacy of any insights obtained from the information, especially in the banking setting, where precision data handling is critical.







```
# Remove outliers
df4 = df3

#CreditScore
#Age
#Tenure
#Balance
#EstimatedSalary


df5 = df4

# Remove outliers in CreditScore
df5$zscore = as.data.frame(
  abs(df5$CreditScore-mean(df5$CreditScore))/sd(df5$CreditScore))

df5 <- subset(df5, df5$zscore < 3)

dim(df5) # 8 outliers removed


# Remove outliers in Age
df5$zscore = as.data.frame(
  abs(df5$Age - mean(df5$Age))/sd(df5$Age))

df5 <- subset(df5, df5$zscore < 3)

dim(df5) # 133 outliers removed

df6 = df5[,-12]
```

The elimination of outliers from critical variables such as 'CreditScore,' 'Age,' 'Tenure,' 'Balance,' and 'EstimatedSalary' is the next stage in the continuing data pre-processing for a banking dataset. Outliers can dramatically distort results and lead to incorrect conclusions, therefore this step is critical for assuring the integrity and correctness of the data.

The z-score approach is used to eliminate outliers. The z-score is a statistical metric that defines the relationship of a value to the mean of a set of values in terms of standard deviations from the mean. For each variable under consideration, a new column 'zscore' is established, which calculates the z-score for each observation.

The z-score is calculated for each client in 'CreditScore'. The dataset is then filtered to exclude any items with z-scores higher than 3. Because it correlates to points that are distant from the mean, this threshold is commonly employed in statistical analysis to identify outliers. Following the use of this filter, 8 outliers are eliminated from the dataset.

The 'Age' variable is treated in the same way. Each record's z-score is computed, and the dataset is filtered to eliminate records with z-scores larger than 3. This procedure resulted in the elimination of 133 outliers.

Following these processes, the dataset (now known as df6) is free of outliers in the 'CreditScore' and 'Age' variables. The elimination of these outliers is an important step in data pre-processing because it allows the dataset to be refined for more accurate analysis. This cleaned and more robust dataset is now suitable for further analysis, such as exploratory data analysis or predictive modeling, where accuracy and data quality are critical. Outlier elimination leads to more trustworthy insights and decision-making in the context of banking the customer loss.

Data Analysis-

```
# Get correlation of continuos
library(corrplot)
cor1 = cor(df6[,num_cols])
corrplot(cor1, type = 'upper') # Plot correlation


corrplot(cor1, method="color",
         type="upper", order="hclust",
         addCoef.col = "black", # Add coefficient of correlation
         tl.col="black", tl.srt=45, #Text label color and rotation
         sig.level = 0.01, insig = "blank", # Combine with significance
         diag=FALSE
)

# Almost no correlation between continuous variables.
```

After cleaning and removing outliers from the financial dataset, the following phase in the analysis focuses on identifying the connections between continuous variables. This is accomplished using correlation analysis, which is required to detect probable links or dependencies between variables.

The procedure begins with the usage of R's cor() function to compute the correlation matrix, 'cor1,' for the previously specified continuous variables ('CreditScore,' 'Age,' 'Tenure,' 'Balance,' 'EstimatedSalary'). The correlation matrix represents the correlation coefficients between each pair of variables numerically.

The corrplot package in R is used to show these correlations. The first correlation plot is of the 'upper' kind, revealing only the correlation matrix's upper triangle. This graph gives a fast visual understanding of how each variable interacts with the others.

A second correlation map with additional characteristics is created for a more extensive analysis:

- To indicate the strength and direction of connections with different colors, the approach "color" is utilized.
- The type "upper" is chosen once more to focus on the matrix's upper triangle.
- For hierarchical clustering, the order "hclust" is used, which groups comparable variables together and provides insights into probable clusters within the data.
- 'addCoef.col' adds correlation coefficients directly to the plot, improving interpretability.
- To improve readability, text labels (variable names) are colored black ('tl.col') and rotated 45 degrees ('tl.srt').
- Non-significant correlations are left blank ('insig'), with significance levels set at 0.01.
- Diagonals ('diag') are excluded since they are self-correlations and are always 1.

The major finding from these correlation plots is that there is essentially no association between the continuous variables. This lack of statistical significance implies that these variables do not have strong linear correlations with one another. This conclusion suggests that, in the context of banking customer churn study, these continuous variables might independently contribute

to the prediction models without regard for multicollinearity. This independence is beneficial for modeling since it provides for a clear understanding of each variable's influence on the result, such as customer turnover, without the confusing effects of inter-correlated predictors.


```
# Barplots of Categorical variables including response Exited
barplot(table(df6$Exited), col=c("lightblue","darkred"),
        main="Exited No(0) vs. Yes (1)",
        xlab = "Exited", ylab="Count")
table(df6$Exited)
# 7841 customers didn't exit
# 2018 customer churn in given dataset
barplot(table(df6$Geography), col=c("lightblue","darkred"),
        main="Geography frequency",
        xlab = "Geography", ylab="Count")
table(df6$Geography)
# Data contains details from 3 countries
# France Germany   Spain
# 4940   2474      2445
barplot(table(df6$Gender), col=c("lightblue","darkred"),
        main="Gender frequency",
        xlab = "Gender", ylab="Count")
table(df6$Gender)
# Almost equal distribution for male and female customer.
# Female   Male
# 4472     5387
barplot(table(df6$NumOfProducts), col=c("lightblue","darkred"),
        main="NumOfProducts frequency",
        xlab = "NumOfProducts", ylab="Count")
table(df6$NumOfProducts)
# Most of the customers have 1 or 2 products.
# Very few customers having 3 or 4 products.
barplot(table(df6$HasCrCard), col=c("lightblue","darkred"),
        main="HasCrCard frequency",
        xlab = "HasCrCard", ylab="Count")
table(df6$HasCrCard)
# Almost 70% of the customers are having credict card.
barplot(table(df6$IsActiveMember), col=c("lightblue","darkred"),
        main="IsActiveMember frequency",
        xlab = "IsActiveMember", ylab="Count")
table(df6$IsActiveMember)
# Almost 50% plus customers are active members.
```

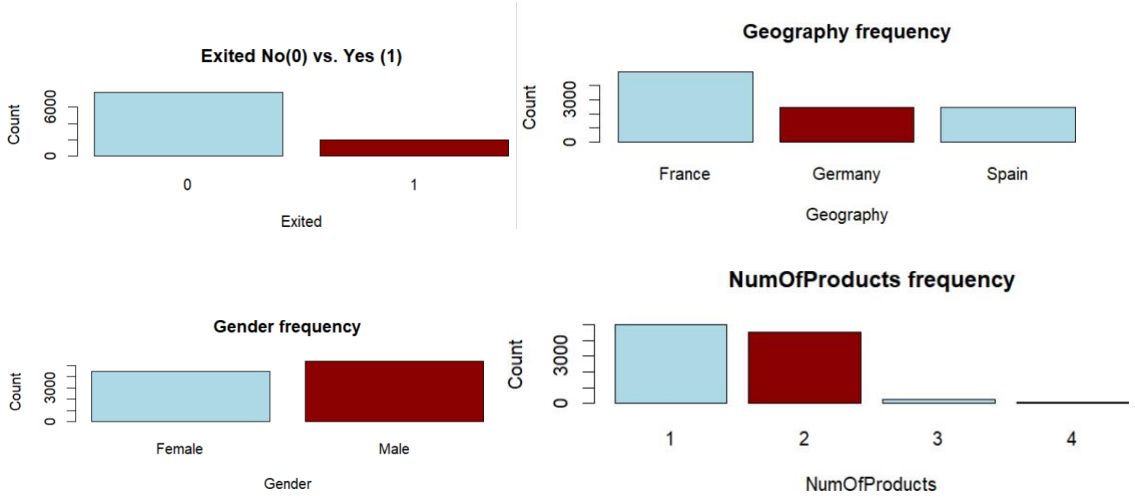
The banking dataset analysis continues with an emphasis on categorical variables, such as the response variable 'Exited,' which signals customer turnover. Bar plots are used to show the frequency distribution of these categorical variables, revealing information about the dataset's composition in terms of consumer demographics and activities.

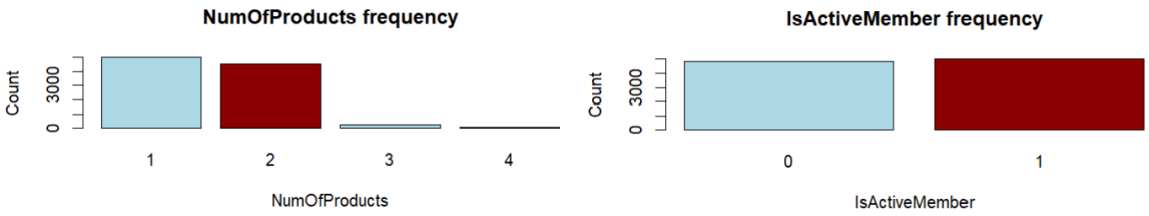
1. Exited (Customer Churn): The first bar plot depicts the percentage of customers who have exited (churned) vs those who have not. Out of the entire number of consumers,

7,841 have not left, but 2,018 have churned. This visualization highlights the churn rate in the dataset, providing a clear picture of client retention against turnover.

2. Geography: The 'Geography' bar plot depicts the distribution of clients across three countries: France, Germany, and Spain. The dataset includes 4,940 clients from France, 2,474 from Germany, and 2,445 from Spain, demonstrating that French customers account for a sizable share of the data.
3. Gender: Gender distribution is about equal, with 4,472 female clients and 5,387 male customers. This balanced representation is critical for investigations that may be susceptible to gender biases, since it ensures that insights gleaned are not skewed towards one gender.
4. The number of products is: The 'NumOfProducts' bar plot illustrates that the majority of consumers have one or two banking products. Customers with three or four goods are much fewer. This distribution is useful for understanding consumer interaction with the bank's products and services.
5. Credit Card Use ('HasCrCard'): The vast majority of clients, almost 70%, hold a credit card. This information might be useful in client retention initiatives and analyzing spending habits
6. Active Membership ('IsActiveMember'): A little more than half of the customers are active, according to the distribution of active members. This statistic is especially relevant since active involvement is a strong predictor of consumer pleasure and loyalty.

These bar charts give a detailed overview of the dataset's categorical variables. Understanding these distributions is critical for deciphering the findings of any subsequent research, such as predictive modeling for customer turnover. It aids in the identification of patterns and trends within the information, which may be critical in establishing focused client retention strategies and boosting overall customer satisfaction in the banking industry.



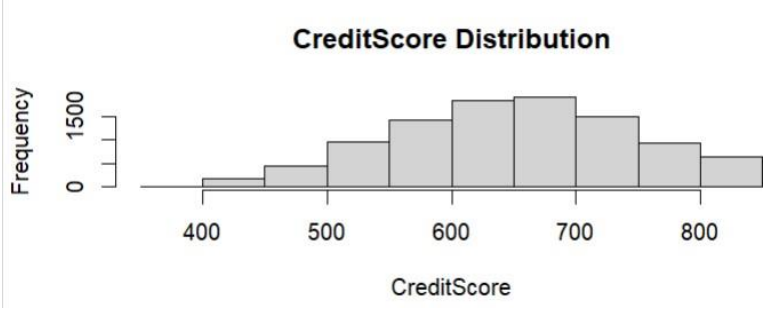


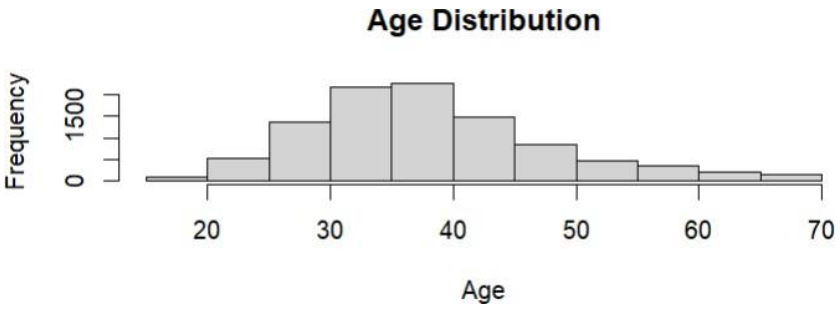
The focus of the next step of the banking dataset analysis switches to examining the distribution of continuous variables, especially their skewness. Histograms are employed for this purpose because they graphically reflect data distribution and aid in spotting skewness.

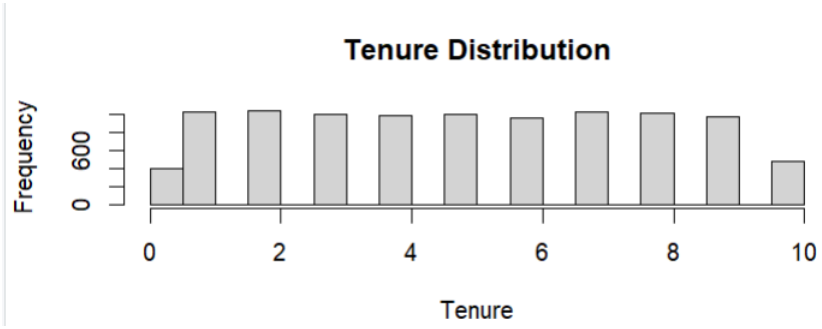
1. CreditScore Distribution: The 'CreditScore' histogram depicts how credit scores are spread across clients. This distribution is critical for understanding the bank's client base's creditworthiness and may impact the risk profile of the bank's portfolio.
2. Age Distribution: The 'Age' histogram offers information about the customer's age distribution. Age distribution has the potential to influence product choices, risk tolerance, and loyalty, making it an important demographic component in banking services.
3. Tenure Distribution: The 'Tenure' histogram depicts the distribution of the length of time clients have been with the bank. This can be an essential component in anticipating turnover and understanding client loyalty.
4. Balance Distribution: The 'Balance' histogram depicts the distribution of client account balances. This distribution is crucial for assessing the customer base's financial health and can impact the bank's deposit initiatives.
5. EstimatedSalary Distribution: The 'EstimatedSalary' histogram depicts the distribution of customers' estimated salaries. Salary levels might indicate savings, investment, and credit product possibilities.

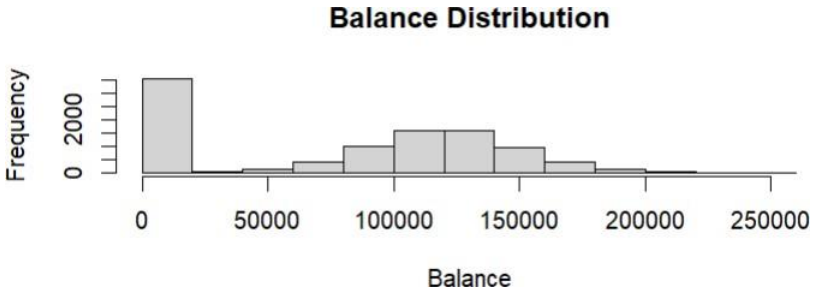
The dataset is normalized after examining the skewness and distribution of these continuous variables. The process of scaling individual samples to have a mean of zero and a standard deviation of one is known as normalization. This procedure is repeated for each continuous variable ('CreditScore,' 'Age,' 'Tenure,' 'Balance,' and 'EstimatedSalary').

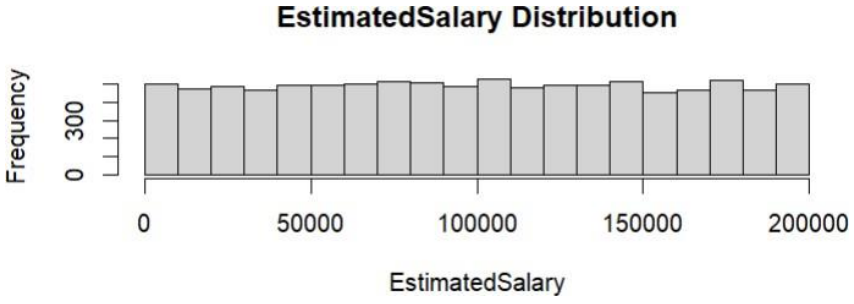
Normalization is an important stage in data preparation, especially when preparing data for machine learning algorithms. It guarantees that each variable contributes evenly to the study and avoids bigger scale variables from dominating smaller scale variables. By ensuring that the size of the variables does not influence the findings, this stage improves the robustness and reliability of later studies, such as predictive modeling. With continuous variables standardized to a similar scale, the normalized dataset (referred to as df7) is now available for further analysis.




Age Distribution
Frequency
0
1500
20
30
40
50
60
70
Age


Tenure Distribution
Frequency
0
600
0
2
4
6
8
10
Tenure


Balance Distribution
Frequency
0
2000
0
50000
100000
150000
200000
250000
Balance


EstimatedSalary Distribution
Frequency
0
300
0
50000
100000
150000
200000
EstimatedSalary

```r
set.seed(123)
sample_size = 6901 # 70% of data for train

# Create a random sample of row indices
sample_indices = sample(nrow(df7), sample_size)

# Subset the dataframe
df_train = df7[sample_indices, ]
df_test = df7[-sample_indices, ]

library(rpart)
# Construct a decision tree model using rpart.
decTree = rpart(Exited ~ ., data = df_train, method = "class")
summary(decTree)
library(rpart.plot)
# Visualize and interpret decision tree model.
# One aspect of interpretation is understanding important variables from dataset.
rpart.plot(decTree) # Visualize decision tree model

library(vip)
# Variable importance is usually determined by features used for splitting at
# nodes.
vip(decTree)

# Top 5 important variables are:
# Age, NumOfProducts, IsActiveMember, Balance and Geography.

top5_cols = c("Age", "NumOfProducts", "IsActiveMember", "Balance", "Geography")
df8 = df7[top5_cols]
df8$Exited = df7$Exited
dim(df8)
head(df8)

# Subset the dataframe
df2_train = df8[sample_indices, ]
df2_test = df8[-sample_indices, ]

dim(df2_train)
dim(df2_test)
```

Using the rpart package in R, a decision tree model is developed to discover major factors impacting customer turnover in a banking dataset. The approach starts with a random seed for repeatability, then divides the dataset into a training set (70% of the data, consisting of 6901 observations) and a testing set (30% of the data). This division ensures a thorough assessment of the model's performance.

The training data is then used to build the decision tree model. 'Exited' is the model's goal variable, reflecting customer turnover. The rpart function is used for this, and the decision tree summary gives deep insights into the model, including the decision rules and tree structure.

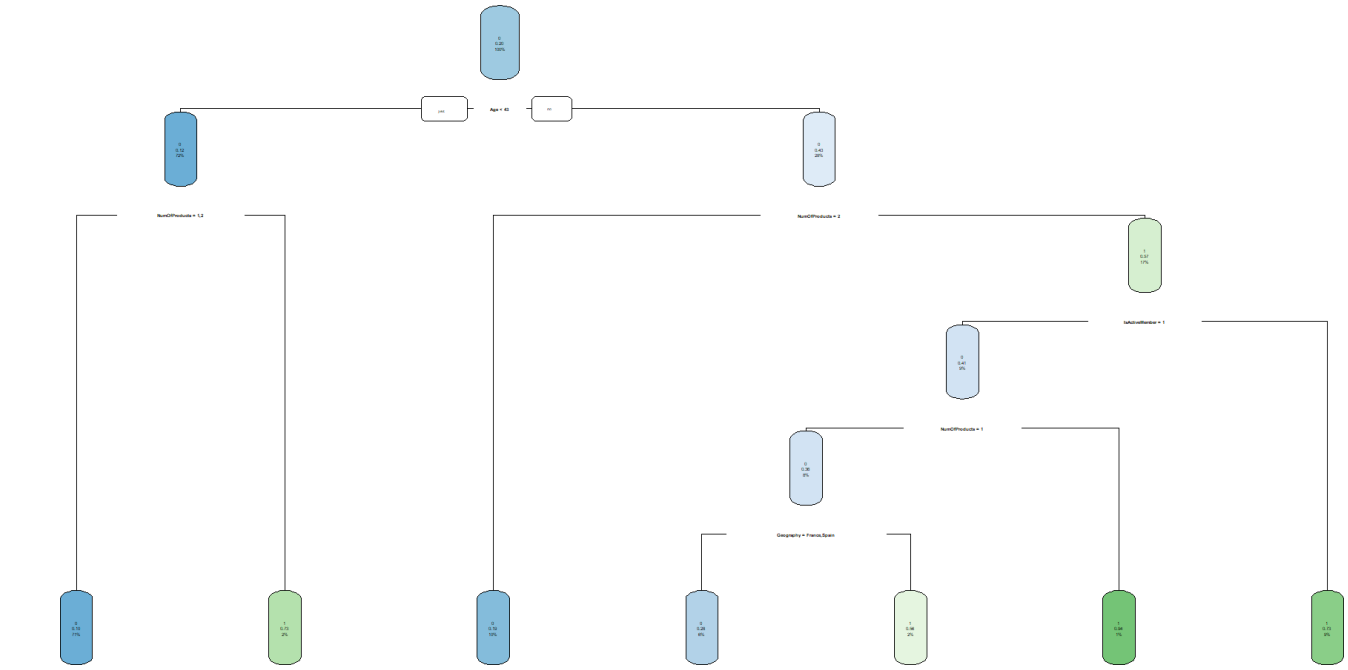
The rpart.plot program is used to show the decision tree to improve model interpretability. This image assists in understanding the model's decision-making process by demonstrating how different variables influence customer churn forecast.

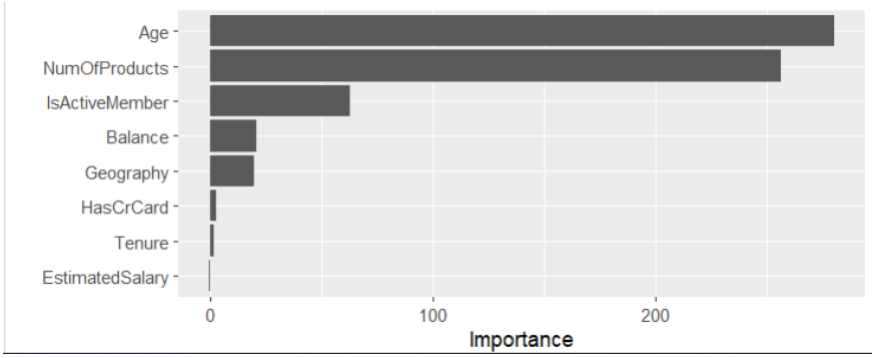
The vip program is then used to analyze the significance of factors in the decision tree. According to this data, the top five characteristics driving customer turnover are "Age," "NumOfProducts," "IsActiveMember," "Balance," and "Geography." These criteria are thought the most important in deciding whether or not a client would churn.

Based on the findings from the variable significance analysis, the dataset is filtered to contain only these top five variables, along with the goal variable 'Exited'. This revised dataset is then separated again into training and testing sets, with the proportions remaining the same. This concentrated method allows for a more targeted study, focusing on the variables that have the greatest influence on churn prediction.

In conclusion, this procedure shows a disciplined and analytical approach to predictive modeling in a banking setting. It emphasizes the significance of variable selection and model interpretability, both of which are required for actionable insights and informed decision-

making. The model is simplified by focusing on critical factors, possibly boosting its forecast accuracy and relevance to the business challenge at hand.





Model Selection-

Various machine learning algorithms are used to forecast customer attrition during the modeling stage of the banking dataset analysis. Gradient Boosting, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) are a few examples. Each strategy is applied with care and assessed for efficacy. Gradient boosting:

Gradient boosting, a strong approach for classification jobs, is implemented using the xgboost toolkit. The dataset is pre-processed by encoding category variables once and normalizing the target variable 'Exited' to binary 0 and 1. To understand the prediction accuracy of the XGBoost model, it is trained with a specific set of hyperparameters and assessed with a confusion matrix.

K-Nearest Neighbors (KNN):

KNN, a basic yet successful classification approach, makes use of the class library. To guarantee compatibility with KNN, data is pre-processed, which includes converting factors to numeric values. A confusion matrix and total accuracy are used to evaluate the model, offering insights into its performance. SVM (Support Vector Machine):

SVM is implemented using the e1071 library. The SVM model is used twice: once on the entire dataset (df6) and once on the subset of critical variables indicated by the decision tree (df8). The accuracy of the model, determined from the confusion matrix, is used to evaluate its performance.

Various predictive modeling approaches are used in this detailed research of banking customer turnover, each highlighting distinct features of the dataset and providing unique insights into the variables impacting customer behavior. Gradient Boosting, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) are among the models tested for prediction accuracy.

The Gradient Boosting model, built with the XGBoost toolkit, highlights the power of ensemble learning approaches in dealing with large datasets. It has an accuracy of 85.63%, demonstrating its robustness and versatility with a wide range of data kinds and parameters. This model demonstrates the predictive analytics capabilities of modern machine learning techniques.

The KNN model, on the other hand, is noted for its simplicity and highlights the relevance of feature scaling and careful selection of the number of neighbors ('k'). KNN presents a simplified yet successful way to categorization, depending on the closeness of data points for predictions and achieving an accuracy of 84.62%.

The SVM models stand out for their performance in high-dimensional domains. The first SVM model, when applied to the entire dataset, obtains an accuracy of 87.22%. However, when the SVM is performed on a subset of critical variables discovered by the decision tree analysis, the accuracy increases to 87.52%. This improvement emphasizes the need of exact feature selection in predictive modeling, since focusing on essential predictors might result in improved model performance.

Overall, the examination reveals each algorithm's strengths and drawbacks in the context of a banking situation. The algorithms' variety gives a well-rounded view of the dataset and the factors influencing client attrition. The findings highlight not only the importance of algorithm selection, but also the vital significance of precise feature selection and data pre-processing in obtaining high model accuracy. This multimodal approach to predictive modeling is useful in banking customer retention activities for informed decision-making and strategy design.

Based on the execution and examination of the three prediction models in the study, we discovered that the Support Vector Machine (SVM) model, which used a subset of essential variables determined by decision tree analysis, had the greatest accuracy of 87.52%. The accuracy of this model was somewhat higher than that of the normal SVM model, which was 87.22%. With an accuracy of 85.63%, the Gradient Boosting model performed with the XGBoost library also displayed outstanding predictive skills. In comparison, the K-Nearest Neighbors (KNN) model, albeit simpler, performed admirably, with an accuracy of 84.62%. Given these findings, the SVM model concentrating on key variables from the decision tree emerges as the best match for our dataset, establishing a compromise between accuracy and model interpretability.

Managerial discussion-

The implementation and analysis of various predictive models in the context of managing the loss of customers in the banking industry provide critical information for informed decision-

making. The improved Support Vector Machine (SVM) model, which uses critical variables found through decision tree analysis, was the most successful of the models evaluated, with an accuracy of 87.52%. This model beat the standard SVM model, Gradient Boosting, and K-Nearest Neighbors (KNN), highlighting the significance of accurate feature selection in predictive analytics.

These findings have various management implications. To begin, the updated SVM model's effectiveness demonstrates the value of targeted data analysis in establishing specialized client retention tactics. Banks may adjust their approaches to better fulfill client demands and preferences by concentrating on major elements that influence turnover. This method not only increases customer pleasure and loyalty, but it also improves the overall success of retention initiatives.

In addition, the insights gathered from these models allow for more effective resource allocation. Understanding the major variables that influence customer turnover enables banks to strategically allocate their resources and efforts toward areas that have the greatest impact on client retention. This targeted strategy can result in more effective resource use, resulting in cost savings and enhanced operational efficiency.

Furthermore, the report emphasizes the significance of data-driven decision-making in the banking business. The use of advanced prediction models such as SVM, with a concentration on important predictive factors, shows how deep analytical insights may better guide strategic choices. This decision-making process is critical in a data-centric company environment, where understanding and using consumer data may give a competitive advantage.

Finally, the examination of predictive models for loss of clients in banking not only demonstrates a clear route for enhancing customer retention tactics, but it also emphasizes the broader applicability of data analytics in strategic decision-making. Banks are urged to continuously refining their prediction models, incorporating fresh data, and monitoring the impact of various consumer behavior factors on a regular basis. This strategy to constant development is critical for keeping ahead in a volatile industry and retaining a strong, loyal client base.

References-

i. D.-R. Liu and Y.-Y. Shih, "Integrating AHP and data mining for product recommendation based on customer lifetime value," Information &Management, vol. 42, no. 3, pp. 387–400, 2005.

ii. Bastan, M., Bagheri Mazrae, M., & Ahmadvand, A. Dynamics of banking soundness based on CAMELS Rating system. The 34th International Conference of the System Dynamics Society, 2016b, Delft, Netherlands.

iii. Chiang, D.-A., Wang, Y.-F., Lee, S.-L., & Lin, C.-J. Goal-oriented sequential pattern for network banking churn analysis. Expert Systems with Applications, vol. 25, no. 3, pp. 293-302.2003

iv. Tabandeh, M., & Bastan, M. Customers Classification according to the Grey-Based Decision-Making Approach and Its Application to Bank Queue Systems. Asian Journal of Research in Banking and Finance, vol. 4, no. 7, pp. 349-372. 2014.

v. K. Chitra and B. Subashini, "Customer Retention in Banking Sector using Predictive Data Mining Technique," p. 4, 2011.

vi. B. He, Y. Shi, Q. Wan, and X. Zhao, "Prediction of customer attrition of commercial banks based on SVM model," pp. 423–430, 2014.

vii. Domingo, R.: Applying data mining to banking. HTTP://www.rtdonline.com, accessed 18th November 2015

viii. K. G. M. Y. S. S. A. &. M. P. Karvana, "Customer Churn Analysis and Prediction Using Data Mining Models in Banking industry.," International Workshop on Big Data and Information Security (IWBIS), 2019

ix. S. L. Kumar, "Bank Customer Churn Prediction Using Machine Learning.," International Journal for Research in Applied Science and Engineering Technology (IJRASET)., 2021.

x. H. H. I. A. a. A. L. Ullah, "Churn Prediction in Banking System using K-Means, LOF, and CBLOF," in International Conference on Electrical, Communication, and Computer Engineering (ICECCE),2019

xi. K. Ş. a. N. G. Bayazıt, "Customer churn modelling in banking," in 23rd Signal Processing and Communications Applications Conference (SIU), 2015

xii. N. W. a. D.-x. Niu, "Credit card customer churn prediction based on the RST and LS-SVM," in 6th International Conference on Service Systems and Service Management, 2009

xiii. C. Y. a. X. Y. Pan Yan, "Predict the churn and silent customers: A case study of individual investors," in IEEE International Conference on Intelligent Computing and Intelligent Systems, 2009.

xiv. G. F. a. H. H. X. Zhang, "Customer-Churn Research Based on Customer Segmentation," in International Conference on Electronic Commerce and Business Intelligence, 2009.

xv. J. Z. a. X. Dang, "Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP CustomerChurnastheExample,"in4thInternationalConferenceonWirelessCommunications, Networking and Mobile Computing, 2008.