Coursework 1
MSCI-562
Author: Risham Hussain
36523348

Purpose of the Report:

An exploratory analysis of customer data from a bank to understand the key factors impacting customer churn rate, employing visualizations for thorough explanations. The focus is retention of customers, rather than acquiring new customers as it is more cost effective.

1. General look at Data and Features

Method: The training and test sets were combined into a complete dataset to be analyzed for an exploratory analysis. The complete data is stored as 'data'. Central tendency statistics were used to get an idea into the shape and correctness of features and check for irregularities such as missing values or outliers.

Results and Discussion: A general look into data indicates that it has no missing values. The features all seem generally correct. Some exceptions exist in Estimated-Salary where salaries lower than 500 exist, as shown in Figure 1 (A). These can be incorrect estimates, obtained with customers who do not prioritize this bank as primary, therefore have a low monthly balance addition, as they may use other bank or accounts as salary accounts. Although this is unusual but not necessarily invalid. This category of exceptionally low salaries can be treated as a customer class holding secondary accounts.

Next the Balance feature has a high volume of accounts with zero balance (~36%), as shown in the Figure 1 (B). This feature is zero-inflated, and indicative of a large proportion of customers that do not actively use their accounts in this bank. This can be problematic, as the effect of this proportion of inactive accounts can spill into other features and upset the statistics obtained from them, for example, which age or tenure categories are reflective of the inactive accounts? Do they impact the members to non-members ratio? Do they inflate the non-members ratio? They may upset other features as well and impact the conclusions drawn from their analysis, adding bias or error. Another example is the "Complaint" feature. If those customers left without filing a complaint, then they inflate the "No complaint" category, giving an illusion of well delivered services by the bank. This can also represent a "Silent Exit" phenomenon, where the customers have stopped using their accounts in this bank but do not formally close it, thus upsetting the exited-to-retained ratio and causing misleading conclusions.

For this, the customers having zero-balance and Inactive status were separated and studied. Their summary statistics did not appear distinct, relative to the total data trend. Since they impart no pronounced individual impact and comprise members that have not yet exited the bank, they are retained for further analysis.

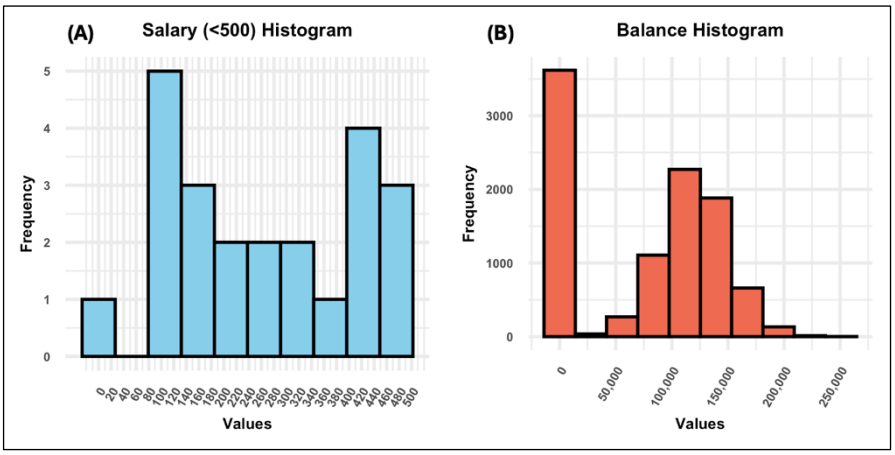

Figure 1Distribution of Salary and Balance features of the dataset, higlhighted zero-inflation and low value accumalation.

2. Comparative Insights from Quantitative Variables: Exited Customers vs Continuing Customers

Method: The customers that have exited the bank were separated into a data frame called 'exited', and the ones continuing bank service into "continuing". Quantitative variables from both datasets were analyzed through data manipulation and boxplots. Additionally conditional boxplots were used for cross comparison.

Results and Discussion: Quantitative variables (Credit score, Age, Tenure, Balance, Estimated Salary and Points Earned) were explored through boxplots for differences. Separating the data frame into the two classes of exited should produce boxplot results similar to conditional boxplots. An observable difference in the boxplot measures between the two datasets would indicate a good feature for discriminating between the exited versus continuing customers. Results indicated that amongst all, 'Age' may be a good classification candidate for customer churn, evident as differences in boxplot measures (quantiles), in the Figure 2.

Additionally, 'Balance' showed differences as well but there exists zero inflation in the data frame of continuing customers. This needs to be dealt with before comparing quantiles as it offsets the values. For this, the zero balance data points were removed, and boxplots were made again. Now the boxplot quantiles show variation across the two data frames, suggesting some discriminatory power in classification. Whether the difference is significant or not, still needs to be determined.

This led to the suggestion that the feature where bank decides whether the customer is an active member is flawed. Many inactive members have large standing balance in their accounts up to ~214,000. One possible explanation could be that the accounts stand as a savings account, however in this case the inactive status wouldn't apply. Amongst the inactive accounts, another discrepancy exists where ~32% of the customers have exited the bank, while having a standing balance up to ~213,000. This is highly unusual as real-life scenarios would dictate that people empty their accounts before exiting, so they don't suffer a loss. Secondly, amongst the active members, ~15% have exited the bank. This does put the validity of the data into question, and puts forth questions such as; (i) does the active status indicate a very active customer that has immediately exited? (ii) Is the active status based on flawed or outdated information? In both cases, it is unreliable, and might not be considered for classification later.

Conclusion: The quantitative variables were analyzed with conditional boxplots. Results showed 'Age' to be a strong discriminator, and "Balance" as a moderate discriminator between the customer churn classes (exited vs continuing customers). However, the Balance is zero inflated and the "Active Member" feature is unreliable and discrepant as it clashes with the "Balance" and "Exited" features. An example is that many inactive members that have exited the bank, have a large standing balance in their account, while many "Active members" have exited the bank. "Active Member" feature is unreliable and might not be considered for classification later.

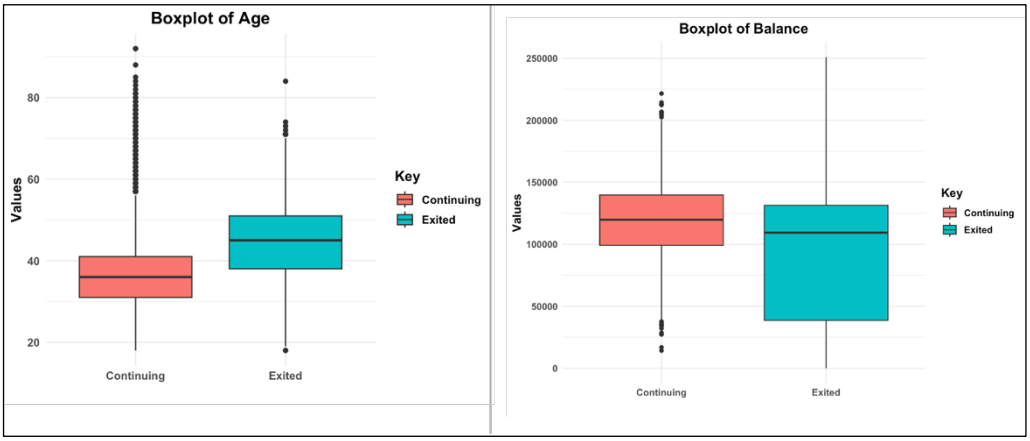

Figure 2 Conditional boxplots of Age and Balance features. The shift in medians of Age and difference in distribution of Balance (shape of boxplot) indicate their ability to distinguishing resultant classes.

3. Factor based Feature Analysis with Conditional and Relative Probabilities:

Method: The features in the dataset that are factors, are analyzed using conditional probabilities to overcome the bias of exited customers, as an imbalanced class.

Results and Discussion: Features that are factors are analyzed using conditional probabilities. This technique is used as the "Exited" label is highly imbalanced in classes (~80% are 'No' and ~20% are 'Yes'). This imbalance of classes will be reflected easily in relative frequencies as it depends upon the total class count. However, conditional probability overcomes the bias of class imbalance and acts as a better determinant for feature selection intended for classification. Features possessing a high conditional probability of the target variable classes, given their values, are considered more important for classification. However, it is important that the feature classes exhibit varying conditional probabilities for different target variable classes, as it is indicative of distinction in classification.

This was put to the test and the categorical features were tested. Amongst all, the features depicting whether customers filed a complaint ('Complain'), are active members ('Is Active Member'), purchased products ('Number of Products'), customer account location ('Geography') and to a lesser extent customer 'Gender' exhibited more importance to be used in classification, The ability of features to separate the resultant classes (Exited or Continuing customers) is evident from the conditional probabilities of their classes, shown in Figure 3.

Additionally, previous results have highlighted flawed or contradictory variables putting the validity of other features into question as well. A second look at the factor data highlighted two more features that contradicted each other; (i) Whether the customer has a credit card and (ii) the type of credit card they have. Logically, if the customer does not have a credit card, the credit card type should be null. However, in this dataset, customers with no credit card still have credit

card types allotted to them (Diamond card:446, Gold:444, Platinum:449 and Silver:452 customers respectively). This can have two explanations, either one of the variables is completely incorrect or some data entry mistake has been made and needs to be amended for a particular class (such as customers with no credit card). Although both the variables do not have a strong predictive power for the resultant customer churn, it was chosen to go with the latter, the data was adjusted for the customers having no credit card to maintain logical sense. Customers with no credited card were allotted a none type variable in card types. With this change incorporated, the card type feature's predictive ability for customer churn remains unchanged.

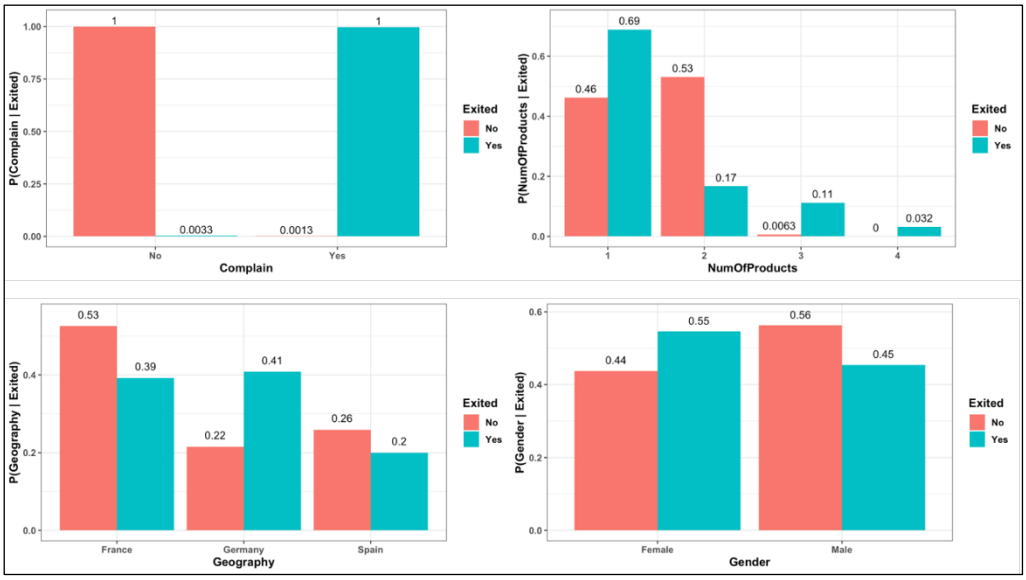

Figure 3 Conditional probabilities with Bar plots of Complaints filed, Number of Products purchased, Geogrpahic location and Gender. These feature classes exhibit varying conditional probabilities for different target variable classes,which is indicative of their ability to distinguish between customer churn classes for classification.

4. Weights of evidence and information value

Method: Inbuilt R functions from "Information" package were used to cross-reference the preceding analyses of categorical and quantitative variables.

Results and Discussion: Information value corroborated what we had concluded in the previous analyses and identified similar features of importance i.e. Customer complaints, Number of Products purchased, Geography, Active member status and gender amongst the categorical variables, Age, and balance from quantitative variable amongst the top choices adding value to the demarcation of customer churn classes. It goes a step further to quantify the extent to which each feature aids in customer churn prediction, while the WOE reveals the prediction or classification strength of each class within the features. In this case, since the customer churn classes are two, the WOE of features can predict the strength of prediction for the class in focus (Exited class in our case), as well as the opposite class by a greater value with the opposite sign. The Information value, however, just reveals the strength of prediction for the classification class in focus, giving a large or small non-negative value.

The Figure 4 illustrates an example of the results. According to the conditional boxplots of Age, the "Continuing" class of customer churn has a median around 35, and its distribution lies below ~40. While the "exited" class has a median of 45, and its distribution lies above ~40. This is reflected in the WOE, as age bons 39 and above provide more information about the class in focus

(exited), while age bins below 39 provide more information about the other class (continuing / not-exited). The more positive or negative the value of WOE, the more predictive strength it has for the respective class. Age bin 18-30 is has the greatest strength of predicting "continuing" while 46-92 predicts "exited" class strongly. Similarly for categorical variable "Number of Products" purchased, conditional probability bar plots indicate that category 2 and 3 are the stronger predictors of the different customer churn classes. This is seen reflected in the WOE plot where category '2' is predictive for "Continuing" class, and category "3" for the class in focus (exited).

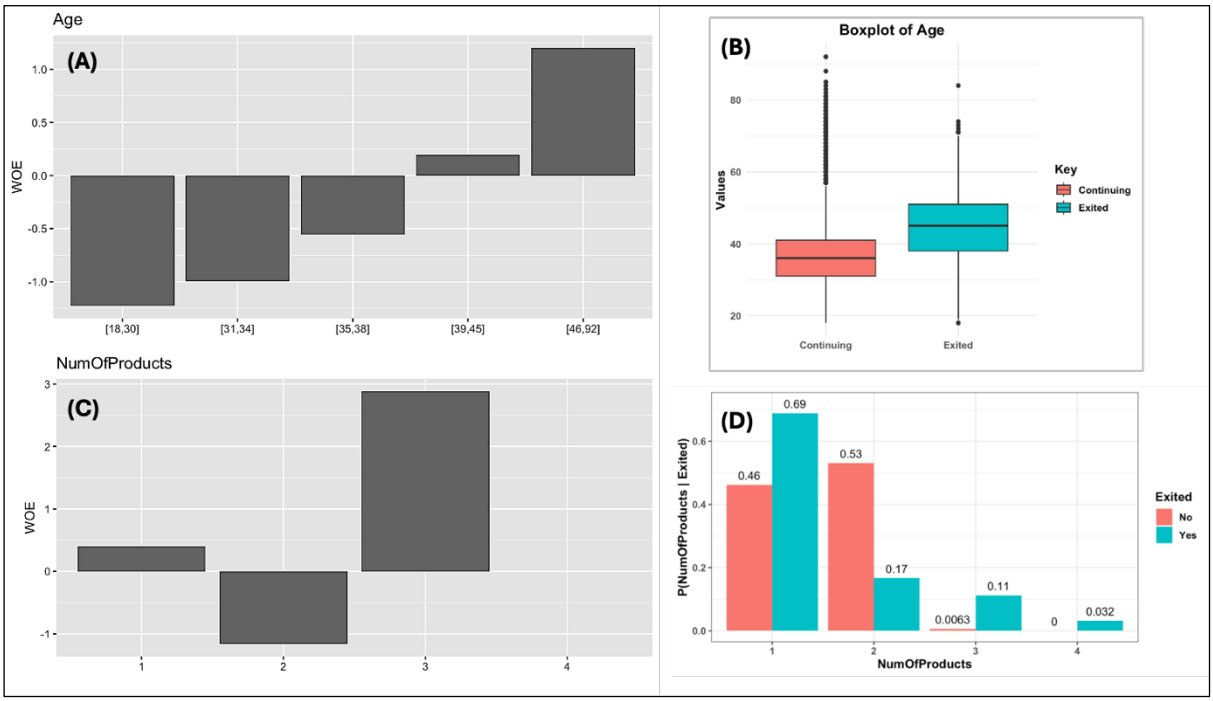

Figure 4 Weight of Evidence plots (left) compared to conditional boxplots (right) for Age and Number of Products fetures, for a comparative analyses.

5. Multi-Dimensional Scaling (MDS)

Method: Inbuilt R functions were used to calculate the distance matrix using Gower's distance as the data has mixed variables. Due to the large dataset taking considerable time to carry out MDS, stress plot was not made, rather three dimensions were used. Stepwise linear regression was used to identify the correlation of the new dimensions with data features.

Results and Discussion: In this analysis, MDS is employed to explore the market segment or the customer segments where the product delivers well, as well as where it fails. Three new dimensions are created. Step-wise linear regression identified the new dimensions to be; (i) Dimension 1 is representative of customers having a credit, (ii) Dimension 2 is representative of Customers who have purchased 2 products, filed complaints and are located in Germany,, and Dimension 2 is representative of Gender. More specifically, Dimension 1 carries no weight in distinguishing between the customer churn classes (exited and continuing customers). Dimension 2 does hold this predictive power, and is influenced by customer complaints and location, that is, customers who haven't filed complaints and belong to Germany, are more likely exit the bank services (Figure 5). Dimension 3 illustrates that female customers are more likely to exit. Thus, it can be concluded that the bank services need to be improved based on Location and Gender, and more focus need to be shifted to the complaints being filed. Improvements such

as better service structures to less popular locations, gender focused incentives and feedback-based improvements can be implemented to retain customers.

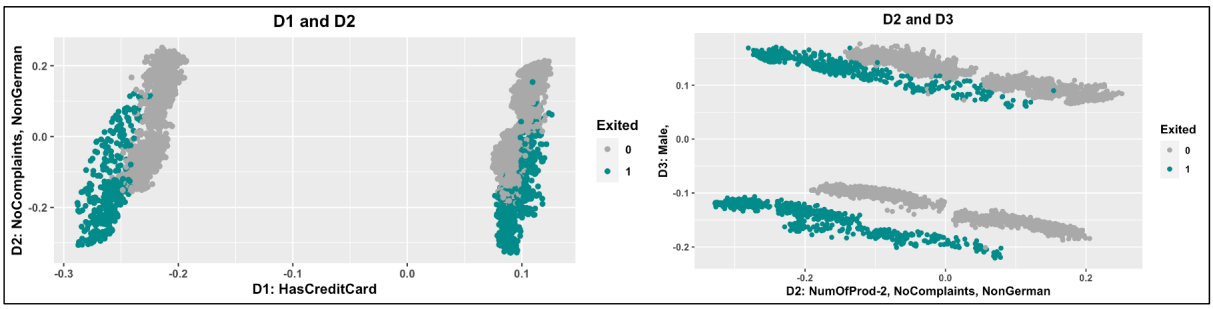

Figure 5 Multi-Dimensional Scaling scatter plots illustrating Dimensions 1 and 2 (left) and Dimensions 2 and 3 (right).

Conclusion

This report presents a detailed exploratory analysis of customer data from a bank, focusing on exploring the dynamics of features and their importance in distinguishing customer churn. The primary aim is to gain an understanding of the factors influencing customer retention, prioritizing retention strategies over acquisition due to their cost-effectiveness. Through quantitative analysis, particularly via boxplots, 'Age' emerged as a robust discriminator between exited and continuing customers, albeit with considerations regarding zero inflation in 'Balance' and reliability concerns with the 'Active Member' attribute. Factor-based analysis, leveraging conditional probabilities, unearthed critical features such as 'Complaint', 'Is Active Member', and 'Geography', pivotal for churn prediction.

Additionally, numerous Irregularities were identified in the data. The “Active Member” feature was discrepant and clashed with the “Balance” and “Exited” features, many inactive members that have exited the bank had a large standing balance in their account, while many “Active members” had exited the bank. The balance feature was zero-inflated, which could be indicative either of customers that do not actively use their accounts in this bank, or a “Silent Exit” phenomenon, where the customers have stopped using their account in this bank but do not formally close it. Further, credit card information features were found contradictory. Customers with no credit card still had credit card types allotted to them, this feature was adjusted to be coherent with possessing a credit card, before continuing with the analysis.

Next, Weight of Evidence and Information values corroborated the preceding results. They strengthened the evidence for features important for distinguishing between customer churn classes by quantifying the impact of identified features on churn prediction, affirming their significance in delineating customer churn classes.

Lastly, multi-dimensional scaling (MDS) unveiled that the limitations in bank services lie in feedback processing of customer complaints, location, and gender-based services. By synthesizing these insights, organizations can tailor retention strategies to bolster long-term customer loyalty and organizational resilience.