1. INTRODUCTION

In the fiercely competitive landscape of the banking industry, retaining existing customers is just as crucial as acquiring new ones. Customer churn, or the rate at which customers discontinue their services, is a key metric that directly impacts a bank's profitability. We delve into customer churn prediction, leveraging a comprehensive dataset encompassing a wide array of customer attributes.

Our primary goal is to build a predictive model that accurately identifies customers who are at risk of churning. We leverage machine learning techniques to provide the bank with actionable insights to engage with potentially at-risk customers proactively.

2. RESEARCH METHODOLOGY

In this paper, we adopt the CRISP-DM (Cross-Industry Standard Process for Data Mining) Framework, widely regarded as the most favored methodology in the data mining domain. By adhering to CRISP-DM, we aim to uncover interesting patterns and insights from the data systematically (Figure 1).

Figure 1: CRISP-DM Approach

Data Understanding

The dataset sourced from Kaggle.com, contains detailed customer information such as credit scores, age, geography, and gender,... Our objective is to utilize this data to train a model and predict the outcome, indicated by the "exit" column, where a value of 1 denotes customer churn. The dataset captures variations based on customer location, economic status, and gender, with the number of products used serving as a proxy for customer loyalty and profitability to the bank. Leveraging such diverse data allows for extracting factual and statistically accurate insights. To facilitate modelling, categorical data were transformed into the numerical format, mitigating information loss. Table 1 provides a summary of the dataset's attributes.

Data Preparation and Visualization

We preprocess the dataset to unify and consistently visualize the diverse input data parameters effectively.

(a) Correlation matrix analysis. Figure 2 depicts the correlation matrix, revealing that no pair of variables exhibits a strong correlation. This observation is crucial as it satisfies the fundamental assumption of modelling, namely the absence of multicollinearity. However, a notable correlation was observed between the number of products and balance.

Table 1: Metadata	
CustomerId	10000
Surname	2932
CreditScore	460
Geography	3
Gender	2
Age	70
Tenure	11
Balance	6382
NumOfProducts	4
HasCrCard	2
IsActiveMember	2
EstimatedSalary	9999
Exited	2
Complain	2
Satisfaction Score	5
Card Type	4
Point Earned	785

Figure 2: Correlation matrix

b) Analysis Based on Balance, Age, Credit Score, and Tenure (Figures 3)

Our investigation reveals that customers who maintain a balance exceeding $85,000 are more likely to churn. Notably, customers aged between 30 and 50 appear to be primary contributors to this trend. Conversely, customers possessing two or more products the bank offers exhibit notably reduced likelihood of leaving. Interestingly, numerical factors such as credit scores and tenure do not significantly impact the customer attrition rate.

Fig 4 showcases a scatterplot illustrating the relationship between credit scores and customer churn. Notably, all churned customers (orange dots) possess credit scores below 400. This suggests that customers with poor credit histories are more prone to leaving the bank. The clustering of churned customers below the credit score threshold of 400 underscores the potential influence of weak economic statuses on customer attrition.

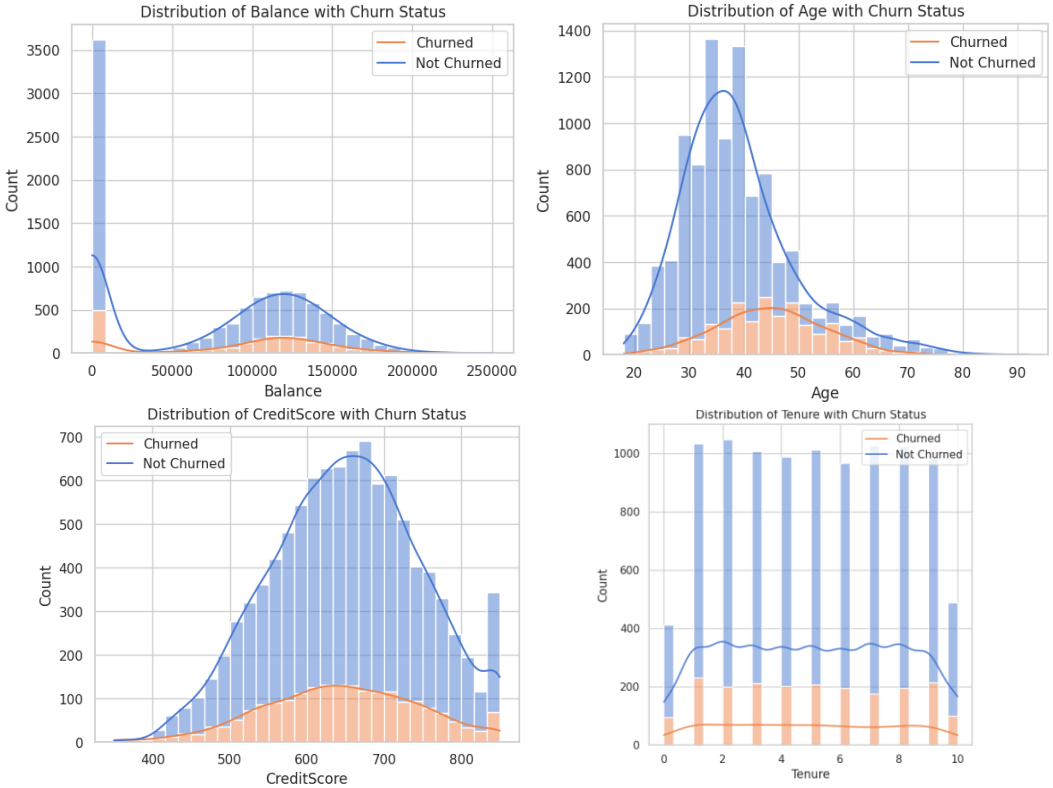

Figure 3: Density plots of (a) observed balance, (b) Age, (c) Credit score, and (d) tenure

(c) Gender, active members, credit cards and country-based analysis. (Figure 5)

We found that 45.4% of female customers churned, while 898 male customers churned out of a total of 5,457 (approximately 16%). Male account for 54.6% of all churning customers. The attrition rate among inactive customers is nearly double that of active customers (27% for inactive compared to 14% for active). Among the countries, France exhibits the highest customer churn rate at 50.1%, followed by Germany at 25.1%, and Spain at 24.8%

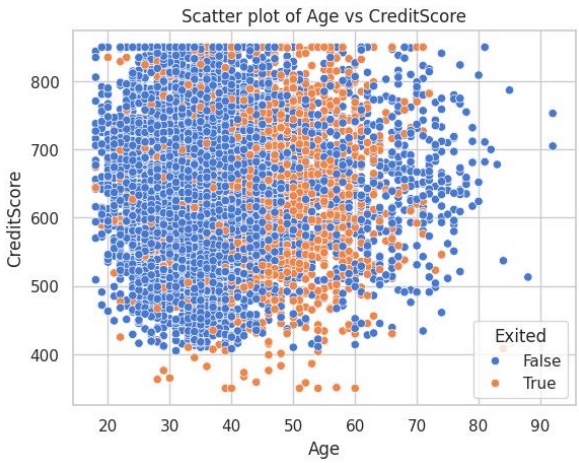

Figure 4: Distribution of customers based on credit score & age

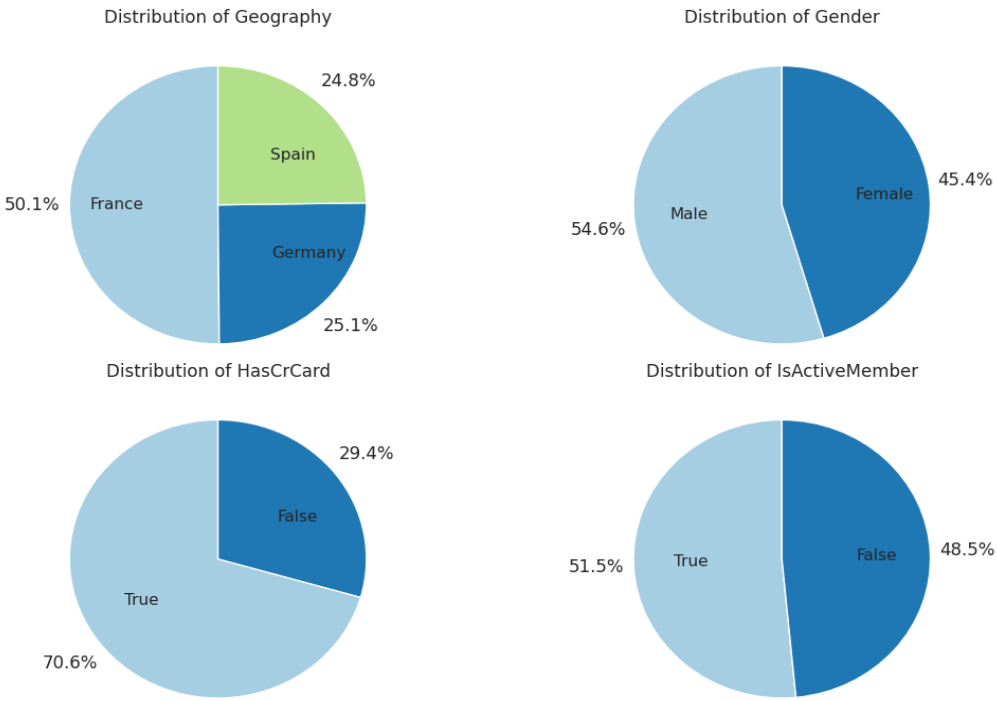

Figure 5: Distribution of Geography, Gender, active members, credit cards

Model Building and Selection

Four pipelines were established, each tailored for logistic regression, random forest, and K-Nearest Neighbors models. These pipelines were applied to a training dataset. To streamline model training, we designed a two-step workflow pipeline.

- Normalize the features to bring them on the same scale
- Instantiate the model to fit

3. RESULTS AND DISCUSSION

Table 2: Evaluation Metrics

Metrics	Logistic Regression	Random Forest	KNN
Accuracy	0.590	0.999	0.711
Specificity	0.475	0.998	0.605
Sensitivity	0.708	1.000	0.819
AUC	0.589	0.999	0.771
F1 Score	0.631	0.999	0.737

The evaluation presented in Table 2 highlights the performance of different models in predicting churn, with a focus on sensitivity and accuracy as the most relevant metrics.

Model Performance:

Random Forest achieved the highest accuracy at 99%, closely followed by KNN at 71.1%. Similarly, Random Forest outperformed KNN in terms of F1 score, specificity, and AUC, indicating its overall superior performance across multiple metrics.

Sensitivity and Accuracy:

Random Forest exhibited the highest sensitivity at 100%, indicating its ability to correctly identify churned customers. KNN followed with a sensitivity of 73.7%.

Since sensitivity and accuracy are considered the most relevant metrics for predicting churn, Random Forest emerges as the preferred choice due to its strong performance in both metrics.

In conclusion, based on the evaluation and considering the importance of sensitivity and accuracy in predicting churn, Random Forest emerges as the preferred choice due to its excellent performance across these key metrics. Random Forest not only demonstrated high sensitivity and accuracy but also excels in handling large datasets with many variables. This scalability makes it well-suited for real-world applications where data volume and complexity are significant.

The analysis identified key factors affecting customer churn, including nationality, gender, number of products owned, and account balance. For instance, German customers showed a higher likelihood of churning than French or Spanish customers, while male customers exhibited a lower probability of churning than females. Moreover, customers with two bank products tend to stay, whereas those with three products are more prone to churning. Customers maintaining a balance exceeding 85,000 are also at a higher risk of churning, possibly due to attractive offers from other banks.

4. CONCLUSION

This study effectively predicts churn among bank customers, offering valuable insights for enhancing customer retention strategies. Banks can leverage these findings to implement tailored approaches aimed at improving customer loyalty and reducing churn rates.

Segmented Offers: Tailor promotions and rewards based on demographics like nationality and gender to appeal to diverse customer preferences.

Product Bundling: Encourage customers to own multiple products by offering bundled packages and incentives for using various banking services.

Proactive Engagement: Reach out to customers with personalized solutions and incentives, especially those maintaining higher balances, to prevent churn.

Retention Incentives: Provide exclusive benefits, such as higher interest rates or waived fees, to retain high-value customers at risk of churning.

Data-Driven Insights: Continuously monitor customer behavior and employ predictive analytics to anticipate churn risk and take proactive measures.

Enhanced Experience: Improve overall customer experience across all channels by streamlining processes and providing personalized service.

Educational Outreach: Educate customers about the benefits of long-term banking relationships through workshops and educational resources.