# Project Title: Understanding and Predicting Customer Churn at Oasis Bank


## Team Members

Rakshanda Hedawoo
Srinivas Varma Chintalapati
Srinidhi Reddy Elkicherla


ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

| Name | Signature |
|---|---|
| Rakshanda Hedawoo | |
| Srinivas Varma Chintalapati | |
| Srinidhi Reddy Elkicherla | |

# Executive Summary:

This research delved into understanding customer attrition within Oasis Bank, aiming to uncover insights crucial for enhancing customer retention strategies and ultimately increasing profitability. By analyzing a rich dataset containing demographic details, banking behaviors, and indicators of satisfaction and churn, we sought to identify the primary drivers behind customer attrition and uncover potential areas for improvement. Our analysis revealed several key findings: Firstly, we discovered that the number of products held by customers significantly influenced churn rates, with those holding three or four products exhibiting the highest churn. Additionally, demographic insights highlighted variations across geographic regions, underscoring the importance of tailored retention initiatives. Financial metrics such as account balance and credit score also played a role in customer churn, suggesting the need for targeted strategies to address disparities across different customer segments. Notably, our analysis revealed that while satisfaction levels were consistent across segments, there was no direct correlation between satisfaction scores and churn rates, indicating the complexity of customer behavior.

What sets our study apart is the comprehensive approach we took in analyzing various facets of customer interaction and behavior within Oasis Bank. By employing advanced analytical techniques and leveraging a diverse dataset, we were able to provide actionable insights that can inform strategic decision-making and drive meaningful improvements in customer retention efforts. Our findings shed light on the intricate dynamics of customer attrition and offer valuable recommendations for Oasis Bank to refine its retention strategies and enhance overall customer experience. This study serves a crucial purpose in helping Oasis Bank stay competitive in a rapidly evolving market landscape by fostering stronger customer relationships and maximizing long-term profitability.

# Data Description:

*Data Source:* Kaggle

*Link*: https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn

*Sample size(n):* 10,000 records

*Number of Variables:* 17 columns

*Sample Data:*

| CustomerId | Surname | CreditSco | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited | Complain | Satisfaction Score | Card Type | Point Earned |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15634602 | Hargrave | 619 | France | Female | 42 | 2 | 0 | 1 | 1 | 1 | 101348.88 | 1 | 1 | 2 | DIAMOND | 464 |
| 15647311 | Hill | 608 | Spain | Female | 41 | 1 | 83807.86 | 1 | 0 | 1 | 112542.58 | 0 | 1 | 3 | DIAMOND | 456 |
| 15619304 | Onio | 502 | France | Female | 42 | 8 | 159660.8 | 3 | 1 | 0 | 113931.57 | 1 | 1 | 3 | DIAMOND | 377 |
| 15701354 | Boni | 699 | France | Female | 39 | 1 | 0 | 2 | 0 | 0 | 93826.63 | 0 | 0 | 5 | GOLD | 350 |
| 15737888 | Mitchell | 850 | Spain | Female | 43 | 2 | 125510.82 | 1 | 1 | 1 | 79084.1 | 0 | 0 | 5 | GOLD | 425 |
| 15574012 | Chu | 645 | Spain | Male | 44 | 8 | 113755.78 | 2 | 1 | 0 | 149756.71 | 1 | 1 | 5 | DIAMOND | 484 |
| 15592531 | Bartlett | 822 | France | Male | 50 | 7 | 0 | 2 | 1 | 1 | 10062.8 | 0 | 0 | 2 | SILVER | 206 |
| 15656148 | Obinna | 376 | Germany | Female | 29 | 4 | 115046.74 | 4 | 1 | 0 | 119346.88 | 1 | 1 | 2 | DIAMOND | 282 |
| 15792365 | He | 501 | France | Male | 44 | 4 | 142051.07 | 2 | 0 | 1 | 74940.5 | 0 | 0 | 3 | GOLD | 251 |
| 15592389 | H? | 684 | France | Male | 27 | 2 | 134603.88 | 1 | 1 | 1 | 71725.73 | 0 | 0 | 3 | GOLD | 342 |
| 15767821 | Bearce | 528 | France | Male | 31 | 6 | 102016.72 | 2 | 0 | 0 | 80181.12 | 0 | 0 | 3 | GOLD | 264 |
| 15737173 | Andrews | 497 | Spain | Male | 24 | 3 | 0 | 2 | 1 | 0 | 76390.01 | 0 | 0 | 3 | GOLD | 249 |
| 15632264 | Kay | 476 | France | Female | 34 | 10 | 0 | 2 | 1 | 0 | 26260.98 | 0 | 0 | 3 | SILVER | 119 |
| 15691483 | Chin | 549 | France | Female | 25 | 5 | 0 | 2 | 0 | 0 | 190857.79 | 0 | 0 | 3 | PLATINUM | 549 |
| 15600882 | Scott | 635 | Spain | Female | 35 | 7 | 0 | 2 | 1 | 1 | 65951.65 | 0 | 0 | 2 | GOLD | 318 |

*Data Description:*

- CustomerId: Unique identifier for each customer.
- Surname: Last name of the customer.
- CreditScore: Numerical variable measuring the credit score of the customer.
- Geography: Categorical variable indicating the country where the customer resides (France, Spain, Germany).
- Gender: Categorical variable indicating the gender of the customer (Male or Female).
- Age: Numerical variable measuring the age of the customer.
- Tenure: Numerical variable measuring the number of years the customer has been with the bank.
- Balance: Numerical variable measuring the account balance of the customer.
- NumOfProducts: Numerical variable indicating the number of bank products the customer uses.
- HasCrCard: Binary variable indicating whether the customer has a credit card (1 for Yes, 0 for No).
- IsActiveMember: Binary variable indicating whether the customer is an active member (1 for Yes, 0 for No).

- EstimatedSalary: Numerical variable measuring the estimated salary of the customer.
- Exited: Binary variable indicating whether the customer has exited the bank (1 for Yes, 0 for No).
- Complain: Binary variable indicating whether the customer has filed a complaint (1 for Yes, 0 for No).
- Satisfaction Score: Numerical variable measuring the satisfaction score of the customer.
- Card Type: Categorical variable indicating the type of credit card the customer holds (DIAMOND, GOLD, SILVER, PLATINUM).
- Point Earned: Numerical variable indicating the points earned by the customer

We'll treat variables like Geography, Gender, HasCrCard, IsActiveMember, Exited, Complain, and Card Type as categorical variables, while the rest will be treated as numerical variables.

# Why this data?

The Oasis Bank dataset provides a comprehensive overview of various aspects of customer interaction, including demographic profiles, banking behaviors, and indicators of satisfaction and churn. By thoroughly analyzing this dataset, Oasis can gain valuable insights into customer preferences, behaviors, and tendencies, which are essential for developing effective retention strategies. Understanding the factors influencing churn, such as customer dissatisfaction or changing financial needs, enables Oasis to proactively address issues, improve service quality, and tailor offerings to better meet customer expectations. Ultimately, leveraging insights from this dataset empowers Oasis to enhance the overall customer experience, foster stronger relationships with clients, and drive long-term profitability and growth.

# Research Questions:

1. What are the main drivers behind customer attrition within Oasis Bank?

2. How do demographic factors such as age, gender, and location influence customer attrition?

3. What impact do financial metrics like credit score, account balance, and income have?

4. How does customer satisfaction, as measured by the satisfaction score, affect the likelihood of attrition?

5. Can predictive modeling accurately forecast which customers are at the highest risk of attrition in the future?

6. What actionable insights can be derived from the analysis to develop targeted retention strategies and loyalty programs?

7. How can Oasis Bank effectively implement and evaluate these strategies to reduce attrition and enhance customer loyalty?

# Methodology:

We employed a structured approach leveraging **PySpark SQL** for initial exploratory data analysis (EDA) to gain insights into the dataset. PySpark SQL provided a convenient interface for performing various operations like **grouping, ordering, filtering, and selecting** data, allowing us to extract meaningful information and understand the underlying patterns. Implemented **PySpark Streaming** to show a summary table that displays aggregated metrics for each geographical region, allowing for a comparative analysis on attrition, complaint count and average satisfaction levels across different areas(Fig.4).

To gain a comprehensive understanding of the customer distribution within the bank's dataset, we utilized **unsupervised learning** techniques such as k-means clustering(Fig.5&Fig.5.1). By applying k-means clustering, we were able to segment customers into distinct clusters based on their attributes, enabling us to assess the attrition rate within each cluster and identify potential trends or patterns.

Subsequently, we transitioned to **supervised learning** techniques including logistic regression, decision trees, and random forest classifiers. These models were chosen for their effectiveness in predicting the likelihood of customer attrition, which is a critical concern for banks. By training and evaluating these models on historical data, we aimed to develop robust predictive models capable of identifying customers at risk of exiting the bank.

Overall, this structured approach enabled us to leverage both exploratory and predictive analytics techniques to gain actionable insights into customer behavior and develop strategies for mitigating attrition within the bank.

# Results & Findings:

**1. Main Drivers of Customer Attrition:**

Churn rates vary with the number of products, with customers having one or two products exhibiting moderate rates and those with three or four products showing significantly higher churn, irrespective of card type(Fig.6).

**2.Demographic Factors:**

Germany has the highest customer count, followed by France and Spain, with a predominance of male customers(Fig.1). Middle-aged customers (26-45 years) are the primary demographic, suggesting potential untapped segments among younger and older age groups.

**3.Financial Metrics' Impact:**

German customers tend to have higher average balances, while credit scores remain consistent across countries(Fig.2). The bank's clientele is segmented into Low-Value High-Risk, High-Value High-Risk, and High-Value Low-Risk groups, highlighting opportunities for risk mitigation and value enhancement strategies.

**4.Customer Satisfaction:**

Satisfaction levels are consistent across different financial segments, with no significant variation between high-value and low-value or high-risk and low-risk customers (Fig.3). However, there's no direct correlation between satisfaction scores and churn rates, indicating the need for deeper analysis.

**5.Behavioral Patterns:**

Clustering analysis reveals clusters with higher churn rates, driven by factors like tenure, age, estimated salary, and complaints. Notably, salary plays a crucial role in distinguishing between these clusters.

**6.Predictive Modeling for Attrition Forecasting:**

Supervised learning models accurately predict attrition risk, with key influencing features including the number of products, age, and complaints. Decision trees outperform other models when main features are removed, while logistic regression excels with all features included.

**Note:** For detailed tables and additional analysis, please refer to the appendix.

# Conclusion:

Our comprehensive analysis for Oasis Bank reveals key insights into customer churn, emphasizing the importance of targeted strategies to enhance customer retention. Through advanced statistical and machine learning methodologies, we identified significant factors influencing churn, including geographical trends, product engagement levels, and customer demographics. Notably, regions like Germany and France exhibit higher churn rates, while specific customer segments, particularly those with multiple product holdings or those expressing dissatisfaction through complaints, are more prone to churn. Our predictive models further distilled these insights, showcasing the paramount importance of product variety, customer age, and feedback responsiveness in forecasting churn likelihood.

Strategic Recommendations:

**1. Customized Engagement Programs:** Implement AI-driven analytics to create personalized customer engagement programs. By analyzing individual customer behavior, preferences, and feedback, develop customized offers, communication, and support that resonate with each customer segment's unique needs, especially targeting high-risk clusters identified in our analysis.

**2. Dynamic Product Alignment:** Develop a dynamic product recommendation system that adjusts offerings based on customer life stages and financial behavior. For instance, younger customers could be engaged with tech-savvy financial tools or investment education, while older segments might appreciate more conservative investment advice or retirement planning services.

**3. Proactive Complaint Resolution Mechanism:** Establish an AI-enhanced monitoring system to identify and address customer complaints before they escalate. Utilize sentiment analysis on customer interactions to detect dissatisfaction and trigger immediate, tailored interventions to resolve issues and communicate the actions taken to the customer proactively.

**4. Geo-Specific Retention Campaigns:** Launch targeted retention campaigns in high-churn regions like Germany and France, leveraging cultural insights and regional customer behavior data. For example, in Germany, where churn and financial engagement are high, introduce exclusive financial advisory services or premium customer support to enhance value perception.

**5. Cross-Functional Customer Success Teams:** Create cross-functional teams combining members from sales, customer service, and product development to provide holistic support and engagement to high-value or at-risk customers. These teams would be responsible for monitoring customer health indicators, developing personalized retention strategies, and ensuring a seamless customer experience across all touchpoints.

By adopting these innovative strategies, Oasis Bank can significantly mitigate customer churn, foster loyalty, and enhance overall customer satisfaction, driving sustainable business growth and competitive advantage in the banking sector.

# Appendix:

| Geography | Gender | count |
|---|---|---|
| France | Female | 2261 |
| France | Male | 2753 |
| Germany | Female | 1193 |
| Germany | Male | 1316 |
| Spain | Female | 1089 |
| Spain | Male | 1388 |

| AgeGroup | count |
|---|---|
| 18-25 | 611 |
| 26-35 | 3542 |
| 36-45 | 3736 |
| 46-55 | 1311 |
| 56-65 | 536 |
| 66+ | 264 |

Fig 1 Demographic factors - PysparkSQL

| Geography | AverageCreditScore | AverageBalance |
|---|---|---|
| France | 649.6683286796969 | 62092.6365157559 |
| Germany | 651.4535671582304 | 119730.11613391782 |
| Spain | 651.3338716188938 | 61818.14776342349 |

| Segment | count |
|---|---|
| Low-Value High-Risk | 3586 |
| Low-Value Low-Risk | 1615 |
| High-Value High-Risk | 3298 |
| High-Value Low-Risk | 1501 |

Fig 2 Financial Metrics - PysparkSQL

| Segment | AvgSatisfaction |
|---|---|
| Low-Value High-Risk | 3.053820412716118 |
| Low-Value Low-Risk | 2.9622291021671825 |
| High-Value High-Risk | 2.9978775015160704 |
| High-Value Low-Risk | 3.0086608927381744 |

| SatisfactionScore | ChurnRate |
|---|---|
| 1 | 0.20031055900621117 |
| 2 | 0.217974180734856 |
| 3 | 0.19637610186092067 |
| 4 | 0.2061752988047809 |
| 5 | 0.19810379241516965 |

Fig 3 Customer Satisfaction- PysparkSQL

| Geography | Exited_Count | Complain_Count | Avg_Satisfaction_Score |
|---|---|---|---|
| Germany | 814 | 819 | 3.0059784774810683 |
| France | 811 | 812 | 3.0177502991623455 |
| Spain | 413 | 413 | 3.013726281792491 |
| Geography | NULL | NULL | NULL |

| Geography | Exited_Count | Complain_Count | Avg_Satisfaction_Score |
|---|---|---|---|
| Germany | 1628 | 1638 | 3.0059784774810683 |
| France | 1622 | 1624 | 3.0177502991623455 |
| Spain | 826 | 826 | 3.013726281792491 |
| Geography | NULL | NULL | NULL |

| Geography | Exited_Count | Complain_Count | Avg_Satisfaction_Score |
|---|---|---|---|
| Germany | 2442 | 2457 | 3.0059784774810683 |
| France | 2433 | 2436 | 3.0177502991623455 |
| Spain | 1239 | 1239 | 3.013726281792491 |
| Geography | NULL | NULL | NULL |

| Geography | Exited_Count | Complain_Count | Avg_Satisfaction_Score |
|---|---|---|---|
| Germany | 3256 | 3276 | 3.0059784774810683 |
| France | 3244 | 3248 | 3.0177502991623455 |
| Spain | 1652 | 1652 | 3.013726281792491 |
| Geography | NULL | NULL | NULL |

Fig 4 PySpark Streaming output

| prediction | ChurnRate |
|---|---|
| 0 | 0.24329501915708812 |
| 3 | 0.23752039151712886 |
| 1 | 0.1498147167813658 |
| 2 | 0.13845350052246605 |

Fig 5 K-means clustering output

| prediction | avg(Tenure) | avg(Complain) | avg(EstimatedSalary) | avg(Age) |
|---|---|---|---|---|
| 3 | 4.961500815660685 | 0.23784665579119085 | 50469.45139641121 | 39.405546492659056 |
| 0 | 5.002554278416348 | 0.24489144316730524 | 149729.90698595162 | 38.98371647509578 |

Fig 5.1 K-means clusters similarity

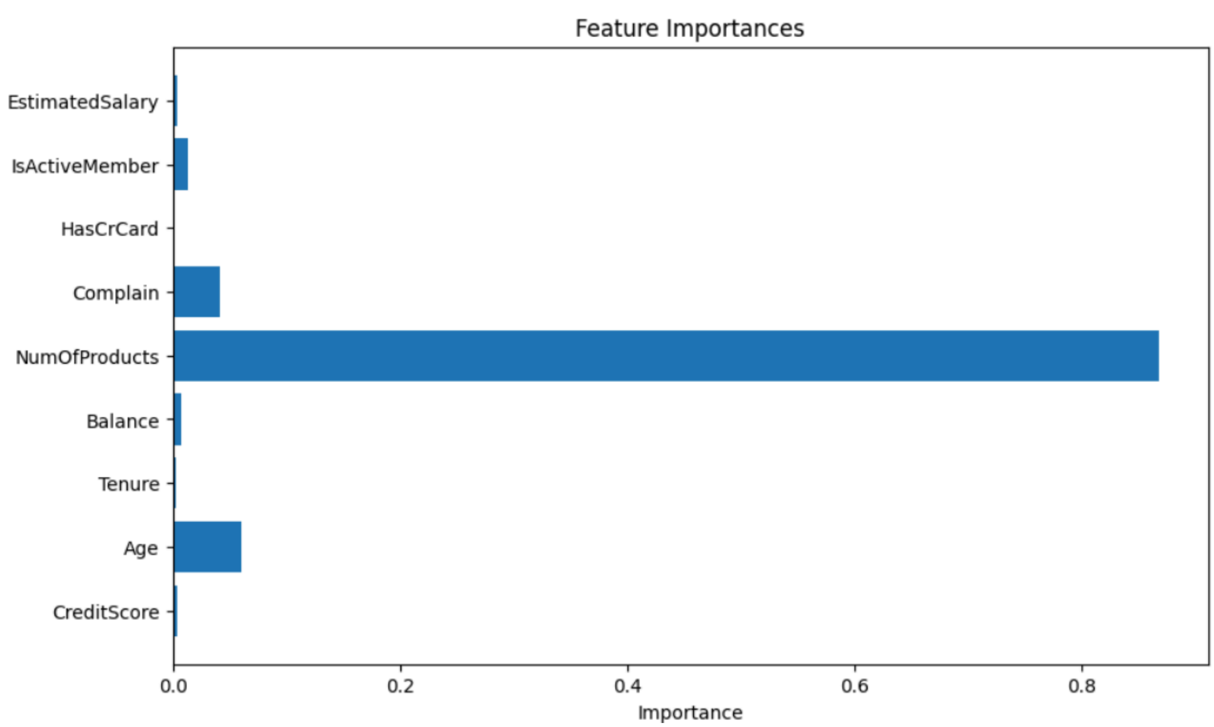


Fig 6. Main Drivers of Customer Attrition