Customer Churn Prediction in Banking

SCTP CAPSTONE PROJECT

PREPARED BY: AMANDA ZHOU

Photo by Etienne Martin on Unsplash

Agenda

- Background
- The Dataset
- Overall Workflow
- Highlights/lowlights
- Model Performance Evaluation
- Key Takeaways & Reflections

Background

Client:

Renown bank that offers various banking products in the highly competitive consumer banking sector. Client operates primarily in France, Germany & Spain.

Current Business Problem:

- The client has been experiencing **increasing** churn rate in recent years.
- High acquisition costs: Customer acquisition can cost up to **5X times more**[1] than customer retention.
- Highly competitive banking sector. Consumers are spoilt for choice. All banks are fighting for the same pie.

Expected Deliverables:

- **What**: Understand the variables that impact the churn rate
- **Why**: To find out why do Customers churn
- **How**: Based on the findings, how the bank can reduce the churn rate
- **Who:** Prediction if a customer will leave
- **What Next**: Recommendations based on findings from this study

1: *Customer Retention Versus Customer Acquisition*

The Dataset

14 Columns, 10,002 records (Numeric & categorical columns)

- **Customer ID:** A unique identifier for each customer
- **Surname:** The customer's surname or last name
- **Credit Score:** A numerical value representing the customer's credit score
- **Geography:** The country where the customer resides (France, Spain or Germany)
- **Gender:** The customer's gender (Male or Female)
- **Age:** The customer's age.

The Dataset – cont'd

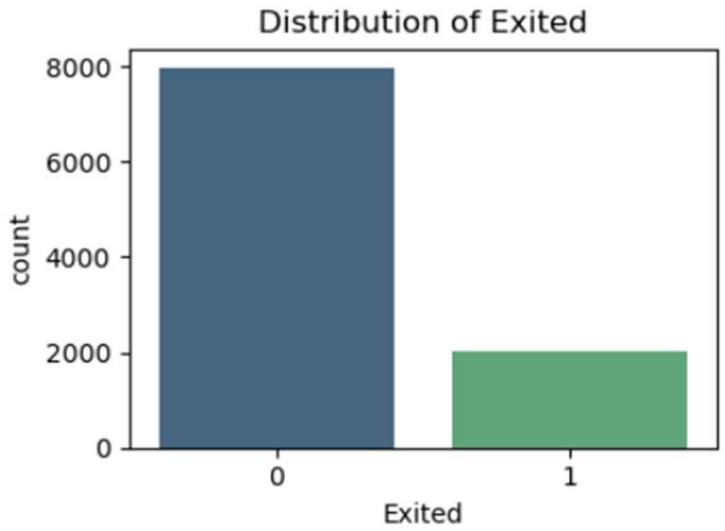
- **Tenure:** The number of years the customer has been with the bank
- **Balance:** The customer's account balance
- **Product Ownership (NumOfProducts):** The number of bank products the customer uses (e.g., savings account, credit card)
- **Credit Card Ownership (HasCrCard):** Whether the customer has a credit card (1 = yes, 0 = no)
- **Active/Inactive Member Status (IsActiveMember):** Whether the customer is an active member (1 = yes, 0 = no)
- **Estimated Salary (EstimatedSalary):** The estimated salary of the customer
- **Customer Churn (Exited):** Whether the customer has churned (1 = yes, 0 = no)

Overall Workflow

Project Workflow – Highlights / Lowlights

Loading & Basic Overview | Visualisation | Exploratory Data Analysis (EDA) | Data Preparation | Feature Selection | Training & Model Performance Evaluation

- Discovered class imbalance in target
  - Not Exited (0) - 79.6%
  - Exited (1) - 20.4%
- Action required:
  - **SMOTE (Synthetic Minority Oversampling Technique) -** creates synthetic examples for the minority class

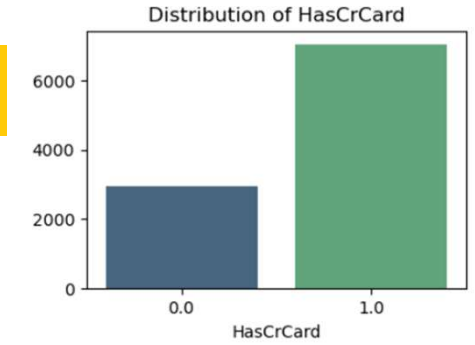


Project Workflow – Highlights / Lowlights

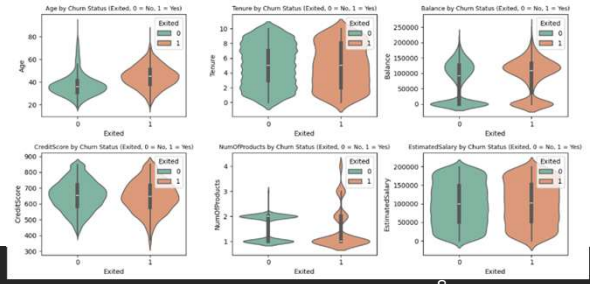
Loading & Basic Overview → **Visualisation** → Exploratory Data Analysis (EDA) → Data Preparation → Feature Selection → Training & Model Performance Evaluation

- Explore Relationships Between Variables – Pairplot
  - **No multicollinearity** – no risk of affecting model performance
- Distribution of Variables
  - **Presence of bimodal** (Balance), **multimodal** data (Product Ownership)
  - **Feature Imbalance** credit card ownership
- Explore Correlation Between Variables and Target
  - Plot numerical variables with violin plot
  - Investigate correlation to decide next course of action for issues identified
  - Credit card ownership >> **no correlation hence no action required**

Feature Imbalance - **HasCrCard**



Violin plot – numerical variables & target



Project Workflow – Highlights / Lowlights

- EDA
  - **Null values** (3 null values) – dropped records as number is insignificant (3 out of 10,002 records)
  - **Redundant columns** (3 columns) – dropped 'RowNumber', 'CustomerId', 'Surname' as they did not hold any significant information

```
# make a copy of the dataframe before dropping the columns
df_cleaned = df.copy()

# drop the rows with null values which are present in the 3 columns
df_cleaned.dropna(subset=['Age', 'HasCrCard', 'IsActiveMember'], inplace=True)
```

```
# drop redundent column - RowNumber, CustomerId & Surname
columns_to_drop = ['RowNumber', 'CustomerId', 'Surname']
df_cleaned.drop( columns = columns_to_drop,inplace=True)
df_cleaned.head()
```

- Data Preparation – **Robust Scaling**
  - Wide & varying range in numerical variables
    - 'Tenure' - 0 to 10 yrs
    - 'Balance' - 0 to 250,000
  - Applied Robust Scaling method to improve model performance, ensuring equal weight of the features.

```
from sklearn.preprocessing import RobustScaler
scaler = RobustScaler()
```

```
# Apply RobustScaler to the dataset
scaled_data = scaler.fit_transform(df_cleaned)

# Convert the scaled data back to a DataFrame for easy viewing
scaled_df = pd.DataFrame(scaled_data, columns=df_cleaned.columns)
```

Project Workflow – Highlights / Lowlights

Loading & Basic Overview → Visualisation → Exploratory Data Analysis (EDA) → Data Preparation → Feature Selection → Training & Model Performance Evaluation

- Model training
  - Applied SMOTE to training data
  - Generates synthetic samples for the minority class (Exited = 1) to balance class distribution

Before SMOTE

```
Original training set class distribution:
Exited
0.0    6350
1.0    1649
Name: count, dtype: int64
```

After SMOTE

```
Resampled training set class distribution (after SMOTE):
Exited
0.0    6350
1.0    6350
Name: count, dtype: int64
```

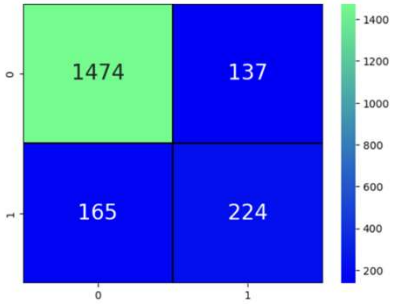
- Models trained:
  - XGBoost
  - Decision Tree
  - Logistic Regression
- Evaluation Metrics:
  - Negative Predictive Value (NPV) – *accuracy in prediction of negative classes*
  - F1-score – *balance in FP & FN*
  - Recall (Sensitivity) – *sensitive to class imbalance, FN is costly*

Model Performance Evaluation

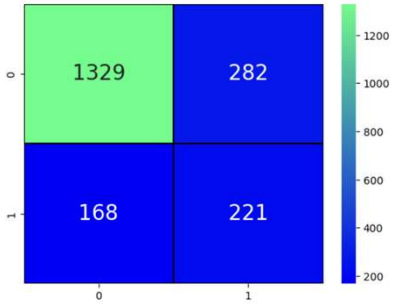
Performance Metrics Comparison:

Model	NPV	F1_score	Sensitivity (Recall)
XGBoost	0.9	0.5973333333333334	0.5758354755784062
Decision Tree	0.89	0.49551569506726456	0.5681233933161953
Logistic Regression	0.92	0.5008880994671403	0.7249357326478149

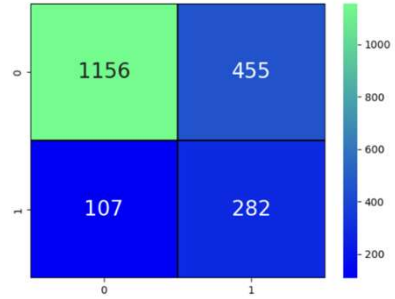
XGBoost:



Decision Tree:



Logistic Regression:



- Best performing model: Logistic Regression
- Strong NPV: 0.92
- High Sensitivity: 0.7249
- Moderate F1-Score: 0.5009 >> Lower than XGBoost but competitive overall.

Key Takeaways & Reflections

Key Takeaways & Recommendations

Customised targeted campaigns by **age group & country**

- Older customers show an increased tendency to leave.
- Differences in churn rates in different countries (Lowest churn rate – Germany)

Attractive **new customer acquisition** campaign

- Sharp dip in number of customers with tenure of 0 years.
  >> fewer new customers in the last 1 year
- Investigate reasons for low onboarding of new customers in the past year. Geography specific?

Further investigation into **product ownership** behaviour

- Low retention rate of customers who own 3 or more products
- Which product(s) they own (Wealth/Investment-related products? Loans?) >> provide insights into banking needs

Investigate correlation between '**Balance**' & '**IsActiveMember**'

- Sizeable portion of the bank's customers hold zero balance in their accounts
- Investigate correlation with member status (Active or Inactive).

Reflections & Areas for Further Analysis

Not all imbalanced features need intervention

Detected imbalanced features in dataset

Rebalancing methods

Unclear definition of parameters

‘Unclear how ‘active’ is measured

Further analyse correlation with other paraments to draw deeper insights.

‘Balance’ – unclear if balance is derived from multiple accounts or account with highest balance.

Success factors in Germany?

Further investigation into lower churn rate in Germany

Macroeconomic factors e.g. Country-specific conditions, consumer behaviour difference

Successful local customer retention campaigns?

More time for further analysis

Further analysis to deep dive into details

e.g. testing correlation between 2 independent variables) in order to better suggest/propose a detailed tactical plan

Thank You!