# Key Results from EDA:

1. **Age and Churn:**
   * Older customers are more likely to exit, with a higher median age among those who churned.
2. **Gender and Churn:**
   * Female customers have a higher churn rate (25.1%) compared to male customers (16.5%).
3. **Geography and Churn:**
   * Customers from Germany have the highest churn rate (32.5%), followed by Spain (16.7%) and France (16.2%).
4. **Balance and Churn:**
   * No significant difference in account balance between exited and non-exited customers.
5. **Credit Score and Churn:**
   * Credit score does not significantly impact the likelihood of customer churn.
6. **Tenure and Churn:**
   * Slightly higher median tenure for exited customers, suggesting tenure has minimal impact on churn.
7. **Number of Products and Churn:**
   * Customers with fewer products are more likely to churn.
8. **Active Membership and Churn:**
   * Active members are less likely to churn compared to inactive members.
9. **Estimated Salary and Churn:**
   * No significant difference in estimated salary between exited and non-exited customers.

**Insight-Based Questions and Answers:**

1. **Which age group is more prone to churn?**
   * Older customers, especially those above 45 years, show a higher tendency to churn.
2. **Is there a gender disparity in customer churn?**
   * Yes, female customers have a higher churn rate compared to male customers.
3. **How does geographic location influence churn?**
   * Customers from Germany are at a higher risk of churning than those from France and Spain.
4. **Does account balance affect customer churn?**
   * No significant impact of account balance on churn rates.
5. **What is the impact of credit score on churn?**
   * Credit score alone is not a strong predictor of customer churn.
6. **How does tenure affect churn?**
   * Tenure shows a slight impact, with longer-tenured customers having a marginally higher churn rate.
7. **Does the number of products influence churn?**
   * Yes, customers with fewer products are more likely to churn.
8. **How does active membership status relate to churn?**
   * Active members have a lower churn rate, indicating the importance of customer engagement.
9. **Is there a correlation between estimated salary and churn?**
   * No significant correlation between estimated salary and customer churn.

# Key result from EDA – revised

### EDA Results Slide (Revised):

#### Key Results from Univariate Analysis:

1. **Credit Score:**
   * Average: 650.53, normal distribution with outliers below 400.
2. **Age:**
   * Mean: 38.92 years, range: 18 to 92 years.
3. **Tenure:**
   * Average: 5 years, range: 0 to 10 years.
4. **Balance:**
   * Mean: 76,485.89, high variation with notable zero balances.
5. **Number of Products:**
   * Majority hold 1 or 2 products.
6. **Has Credit Card:**
   * 71% possess a credit card.
7. **Is Active Member:**
   * 52% are active members.
8. **Estimated Salary:**
   * Average: 100,090.24, wide range of income levels.
9. **Exited:**
   * 24% have exited.

#### Key Results from Bivariate Analysis:

1. **Age and Churn:**
   * Higher churn among older customers.
2. **Gender and Churn:**
   * Higher churn rate in female customers.
3. **Geography and Churn:**
   * Highest churn in Germany.
4. **Balance and Churn:**
   * Balance does not significantly impact churn.
5. **Credit Score and Churn:**
   * Similar credit score distribution in both groups.
6. **Tenure and Churn:**
   * Slightly higher churn with longer tenures.
7. **Number of Products and Churn:**
   * Higher churn with fewer products.
8. **Active Membership and Churn:**
   * Lower churn among active members.
9. **Estimated Salary and Churn:**
   * No significant salary difference impacting churn.

# Data Preprocessing

### Data Preprocessing

1. **Duplicate Value Check:**
   * Checked for duplicate rows in the dataset.
   * Removed any duplicate entries to ensure data integrity.
2. **Missing Value Treatment:**
   * Identified columns with missing values.
   * Applied suitable imputation techniques (mean/median/mode) to fill missing values.
3. **Outlier Check (Treatment if Needed):**
   * Detected outliers using the Interquartile Range (IQR) method.
   * Applied transformations or capping to handle outliers where necessary to prevent skewed model performance.
4. **Feature Engineering:**
   * Created new features based on existing ones to enhance model performance.
   * Converted categorical variables into numerical format using techniques like one-hot encoding.
5. **Data Preparation for Modeling:**
   * Standardized numerical features to ensure uniform scaling.
   * Split the dataset into training and validation sets for model training and evaluation.

# Neural Network with SGD Optimizer

### Key Performance Metrics

* **Recall on Training Set**: 0.08 - Indicates challenges in correctly identifying exited customers
* **Recall on Validation Set**: 0.07 - Consistent with training recall, highlighting issues in model generalization
* **Accuracy**:
  + Training: 81%
  + Validation: 79%

### Observations

* Despite high accuracy, the model struggles with recall, especially in identifying the positive class effectively.
* The model shows consistent performance across training and validation sets but could benefit from further tuning to address the recall issues.

# Neural Network with Adam Optimizer

### Performance Overview

* **Recall**:
  + Training: 0.67 - Shows challenges in identifying positive classes
  + Validation: 0.57 - Suggests potential issues in model generalization
* **Accuracy**:
  + Training: 91%
  + Validation: 84%

### Observations

* The model demonstrates strong accuracy but experiences a drop in recall, suggesting issues with identifying the positive class, particularly in validation scenarios.
* This indicates the need for further tuning, potentially exploring different architectures or training procedures to enhance recall.

# Neural Network with Adam Optimizer and Dropout

**Performance Summary:**

* **Recall on Training Set:** 0.59 (indicates moderate sensitivity to positive class)
* **Recall on Validation Set:** 0.52 (shows performance consistency with a need for improvement)
* **Accuracy on Training Set:** 0.90
* **Accuracy on Validation Set:** 0.86

**Observations:**

* The model shows consistent accuracy, but the recall indicates some challenges in effectively capturing the positive class, particularly in the validation phase. Dropout layers help mitigate overfitting, maintaining a robust performance across different data samples.

# NN with Balanced Data (by applying SMOTE) & SGD Optimizer

**Optimizer and Training Parameters:**

* **Optimizer:** SGD
  + **Learning Rate:** 0.001 (ensures stable convergence)
* **Metrics:** Recall (to maximize the correct identification of the minority class)
* **Loss Function:** Binary Crossentropy (ideal for binary classification tasks)
* **Batch Size:** 32 (to manage computational resources effectively)
* **Epochs:** 100 (to provide sufficient iterations for learning from the balanced dataset)

**Performance Summary:**

* **Recall on Training Set:** 0.73 (demonstrates the model's sensitivity to the positive class)
* **Recall on Validation Set:** 0.72 (maintains performance on unseen data)
* **Accuracy on Training Set:** 0.74
* **Accuracy on Validation Set:** 0.74 (consistent accuracy across both datasets)

**Observations:**

* The use of SMOTE effectively balances the class distribution, enhancing the model’s ability to identify exited customers.
* The consistent recall and accuracy scores across the training and validation datasets illustrate a stable model that generalizes well, avoiding the common pitfalls of overfitting.
* The chosen learning rate and batch size ensure that the model learns effectively without missing nuances in the data.

This overview encompasses the architecture and configuration of the model, the rationale behind the chosen parameters, and a concise summary of the model's performance, tailored for clarity and impact in a presentation setting.

# NN with Balanced Data (by applying SMOTE) & Adam Optimizer

**Performance Metrics:**

* **Primary Metric:** Recall
* **Training Recall:** 0.95
* **Validation Recall:** 0.66
* **Training Accuracy:** 0.94
* **Validation Accuracy:** 0.80

**5. Training Details:**

* **Epochs:** 100
* **Batch Size:** 32
* **Validation Data Used:** Yes

**6. Observations:**

* The model shows strong performance in identifying exited customers during training.
* However, the performance drops on the validation set, indicating possible overfitting.
* Strategies to reduce overfitting and enhance model generalization might be needed.

# Neural Network with Balanced Data (by applying SMOTE), Adam Optimizer, and Dropout

**erformance Metrics:**

* **Primary Metric:** Recall
* **Training Recall:** 0.86, reflecting high sensitivity in identifying positive class
* **Validation Recall:** 0.75, good but lower than training, indicating some overfitting
* **Training Accuracy:** 0.85
* **Validation Accuracy:** 0.78

**5. Training Details:**

* **Epochs:** 100, for sufficient model training iterations
* **Batch Size:** 32, optimizing computational efficiency

**6. Observations:**

* The model effectively identifies exited customers, particularly in the training set.
* There is a decrease in accuracy from training to validation sets, suggesting the model may be overfitting and could benefit from further tuning or additional dropout layers.