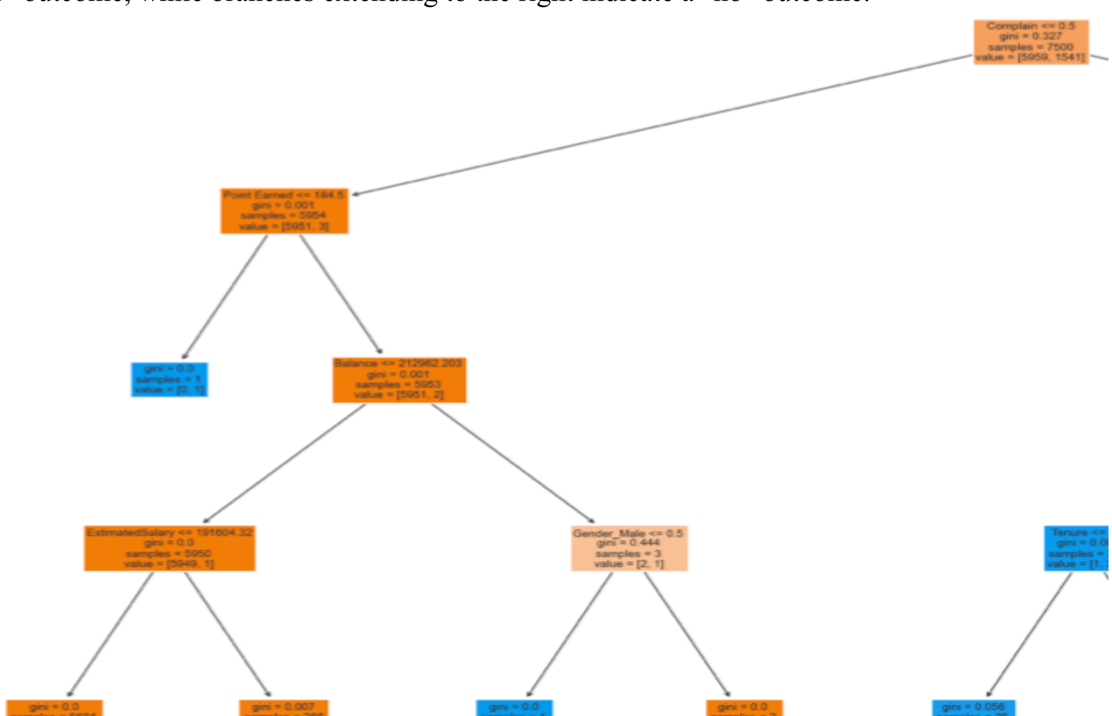
BANK X Churn by DecisionTree

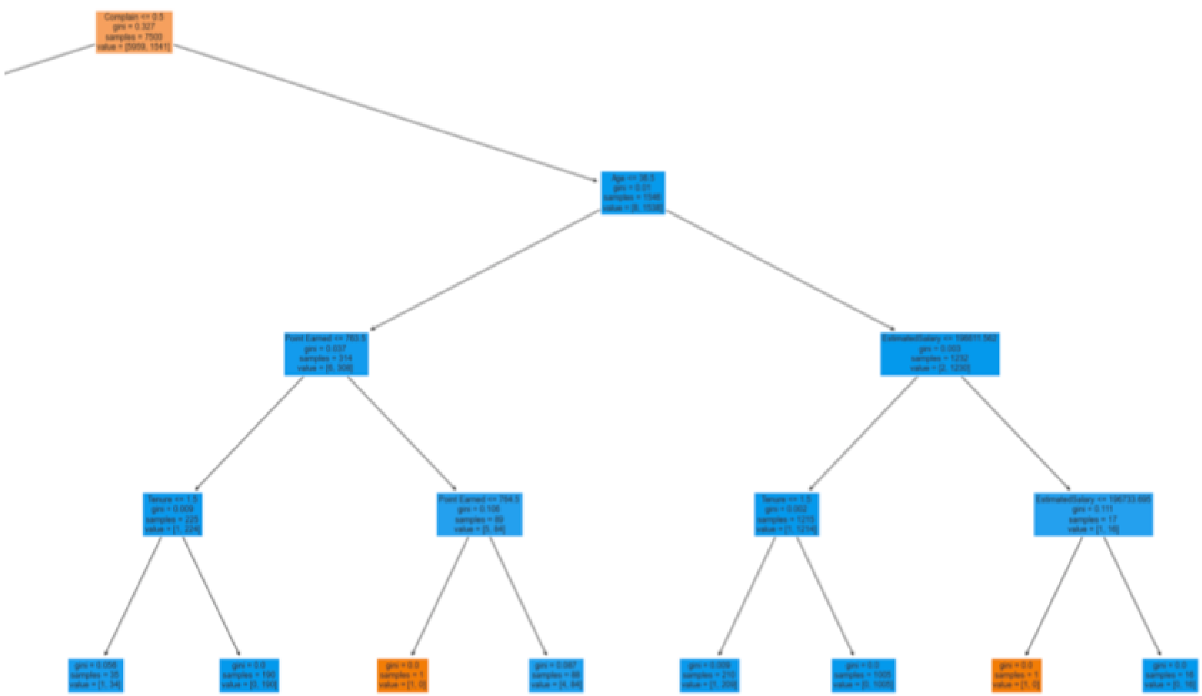
Bank X has encountered difficulties in retaining a significant portion of its customer base. To address this issue, we employed the Decision Tree learning model to identify areas within the company that may be lacking. Decision Trees offer an effective visualization of the impact each variable has on the overall outcome. By examining the gini index at each branch, we can predict the likelihood of customer churn at various decision points.

To initiate our data science project, we commenced with data cleaning. The dataset was obtained from kaggle.com and comprised 10,000 rows with seventeen separate columns. Fortunately, there were no missing values, eliminating the need for imputation or substitutions. We divided the dataset into X and Y fields, excluding the row number column as it had no impact on our analysis. The train_test_split function was then applied to fit our data into the Decision Tree Model.

Upon initial inspection, the Decision Tree appears complex. However, we can dissect the branches and variables to comprehend their significance as we trace our way back to the root nodes. The primary leaf, "Complain," stands out as the variable with the greatest influence on whether a customer will leave the bank. The Decision Tree algorithm iteratively assesses the influence of each initial variable on the data. When it identifies the variable with the best parameter score, typically measured by the gini index, a new branch is created to identify subsequent important variables. Branches extending to the left indicate a "yes" outcome, while branches extending to the right indicate a "no" outcome.

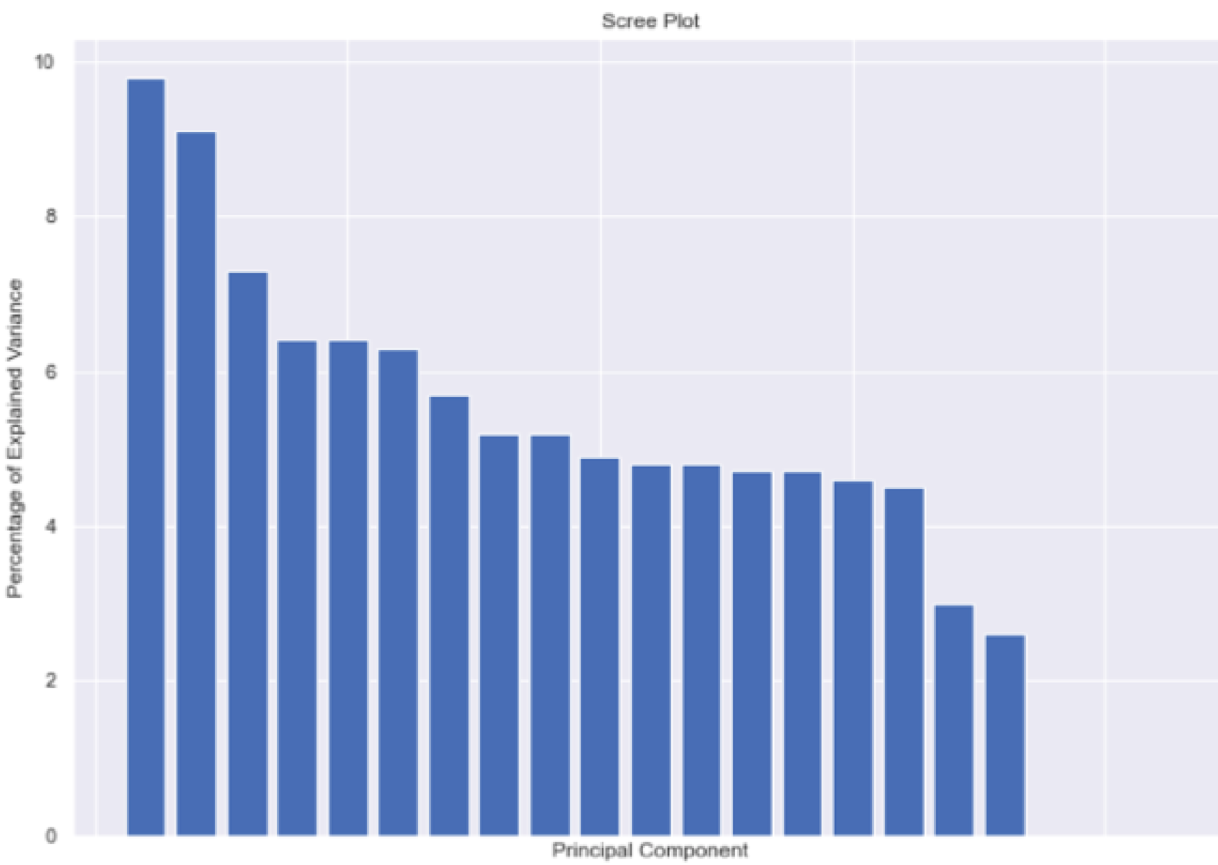


The left half of the Decision Tree, represented by the branch "Complain <= 0.5" (indicating that the customer did not complain), demonstrates that Bank X has relatively little difficulty retaining long-term customers, particularly older men. Customers who have invested more with the bank, as indicated by higher points and balances, are less likely to exit. Blue roots signify customers who have left the bank, while orange ones represent those who have stayed. This left side of the tree predominantly consists of customers who remained with Bank X due to the significant influence of the complaint variable.



On the other hand, the right side of the Decision Tree represents customers who have exited the company. After the initial split based on complaints, the next branch extends to the age variable. Customers under the age of 36.5 exhibit a higher propensity for churn. While the exact reasons behind this trend can only be speculated upon, it is likely that younger customers have less tolerance for inadequate service. The initial complaint leaf indicates that customers in this branch experienced some issue with the bank, and further exploration reveals potential factors contributing to their dissatisfaction, such as low point balances, tenure, and actual balances. It is crucial to recognize the importance of not only rewarding higher-tier members but also addressing the needs of customers with smaller accounts. Although younger customers may currently have low balances, their future potential and likelihood of continuing business with the bank are significantly impacted by their early experiences. This highlights a clear deficiency within the customer service department and its procedures, necessitating a need for change.

Following our Decision Tree analysis, it is essential to validate the influence of the identified variables. Principal Component Analysis (PCA) serves as a valuable tool for visualizing how each variable influences others and contributes to the overall outcomes.



The scree plot above illustrates the amount of variance explained by each Principal Component. We focus on the first PC as it exhibits the highest explained variance. Although the data points do not share significant commonality overall, examining the loading scores of the first component can provide valuable insights.

```
Variable importance ranked 1-20: Gender_Male           0.635391
Gender_Female          0.635391
Complain               0.250608
Geography_Germany      0.226514
Balance                0.155625
Geography_France       0.148253
Age                    0.119022
IsActiveMember         0.065768
Card Type_GOLD         0.064549
Geography_Spain        0.055408
Card Type_DIAMOND      0.051797
NumOfProducts          0.041388
EstimatedSalary        0.027564
Tenure                 0.025781
Point Earned           0.014763
Satisfaction Score     0.014581
Card Type_PLATINUM     0.010968
CustomerId             0.009594
HasCrCard              0.007089
Card Type_SILVER       0.001625
dtype: float64
```

Remarkably, the top variables in PCA—Gender, Complaint, Balance, and Age—coincide with the branches observed in our Decision Tree. This correspondence indicates that the Decision Tree accurately selected the primary variables. While Decision Trees are inherently greedy algorithms that do not always split optimally, the significant influence of complaints on the final outcome lends confidence to the accuracy of our Decision Tree splits.

In conclusion, the Decision Tree analysis for Bank X highlighted the critical role of customer complaints in determining customer churn. The findings suggest a need for substantial improvements within the customer service department to address customer dissatisfaction and enhance customer retention. By focusing on addressing complaints, improving service quality for younger customers, and recognizing the value of all customer segments, Bank X can work towards improving customer retention rates and overall satisfaction.