Aalto University

PROJECT REPORT

Topic: Predicting churn rate among customers of a European bank.

1. Business Overview

In the banking industry, customers usually have more than enough choices regarding where to keep their hard-earned money. This means acquiring customers is only as important, or even less important, than keeping current ones as loyalty is crucial in the financial services industry with client lifetime value being an important indicator.

A client of ours - an anonymous European bank serving France, Germany, and Spain markets - currently has a roughly **20%** of churn rate which is almost double that of the industry average of **11%**, according to a report by ***Accenture***. Acting as their consultants, we are tasked with a simple mission: **how to predict the characteristics of customers who are likely to churn in order to take targeted measures to those specific groups.**

Our plan is to build three predictive models (Decision Tree, Support Vector Machine, and Logistic Regression) to predict if a customer with specific characteristics is more likely to churn. These models will be compared based on ROC, AUC, Accuracy score and Confusion Matrix in order to identify the most accurate model.

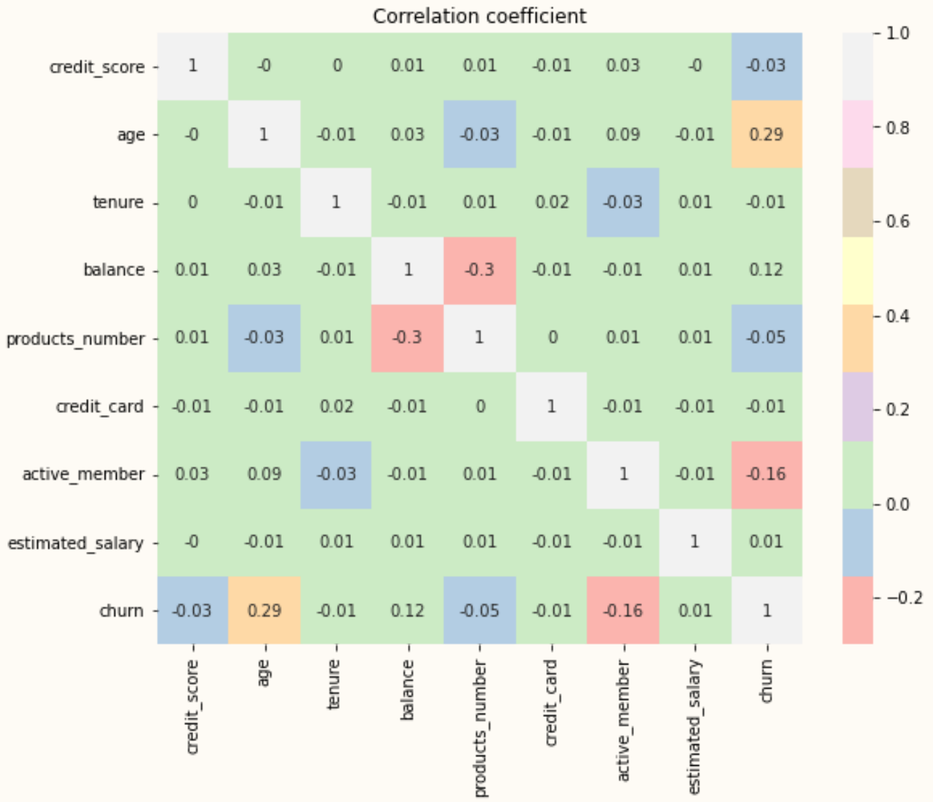
Source: https://thefinancialbrand.com/news/bank-marketing/banking-customer-acquisition-attrition-growth-strategy-68371/

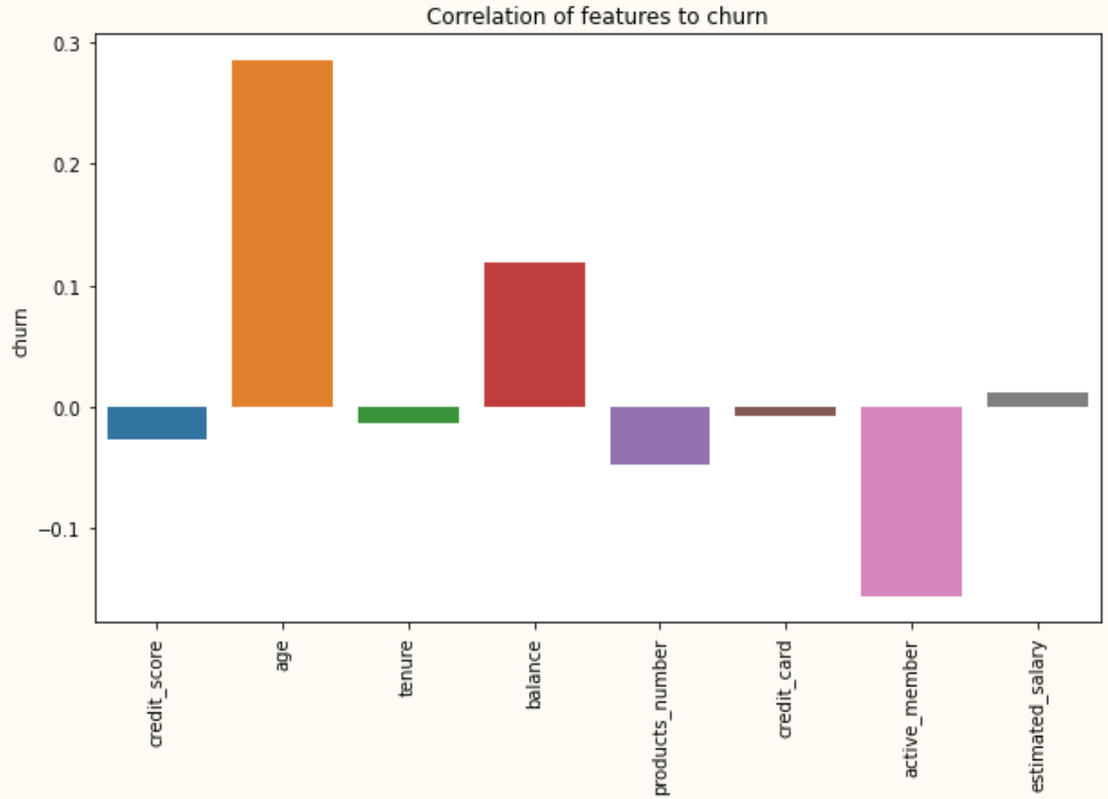
2. Data Understanding – variables

Variable	Description	Values	Type	Action
customer_id	Account number	Distinct 8-digit numbers	Continuous	Remove
credit_score	Credit score of customer	350 - 850	Continuous	-
country	Country of Residence	France, Germany, Spain	Categorical	Change to Boolean
gender	Gender of customer	Female, Male	Categorical	Change to Boolean
age	Age of customer	18 - 92	Continuous	-
tenure	For how many years the customer has had the bank account in ABC bank	0 - 10	Continuous	-
balance	Customer's account balance	0 - 215k	Continuous	-
products_number	Number of products from the bank	1 - 4	Categorical	Change to Boolean
credit_card	Whether the customer has credit card or not	0, 1	Boolean	-
active_member	Whether the customer is an active member or not	0, 1	Boolean	-
estimated_salary	Salary of the account holder	0 - 200k	Continuous	-
churn	Whether the customer churns or not	0, 1	Boolean	Target variable

- **customer_id**: The variable is used by the bank as identifiers. Thus, it has no explanatory value and should be removed.
- **country, gender, products_number**: As these are categorical variables, we converted them to columns with a 1 or 0 by using '**one hot encoding**' approach.
- No Null values in our data, no need for action

2. Data Understanding – correlations

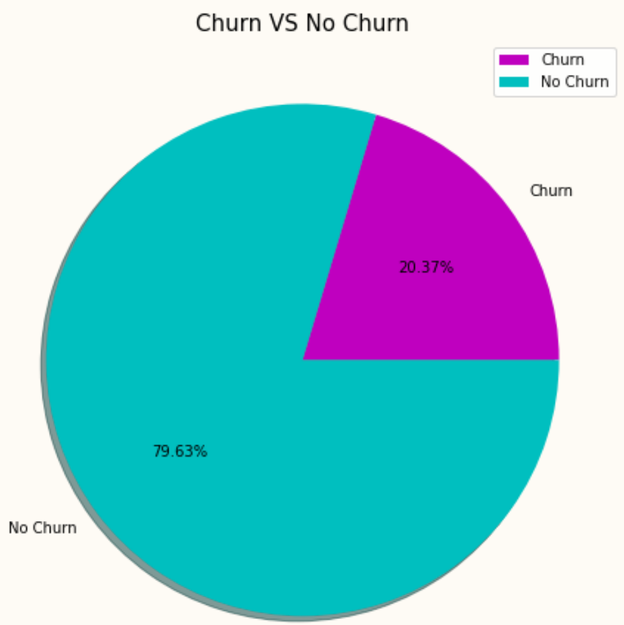


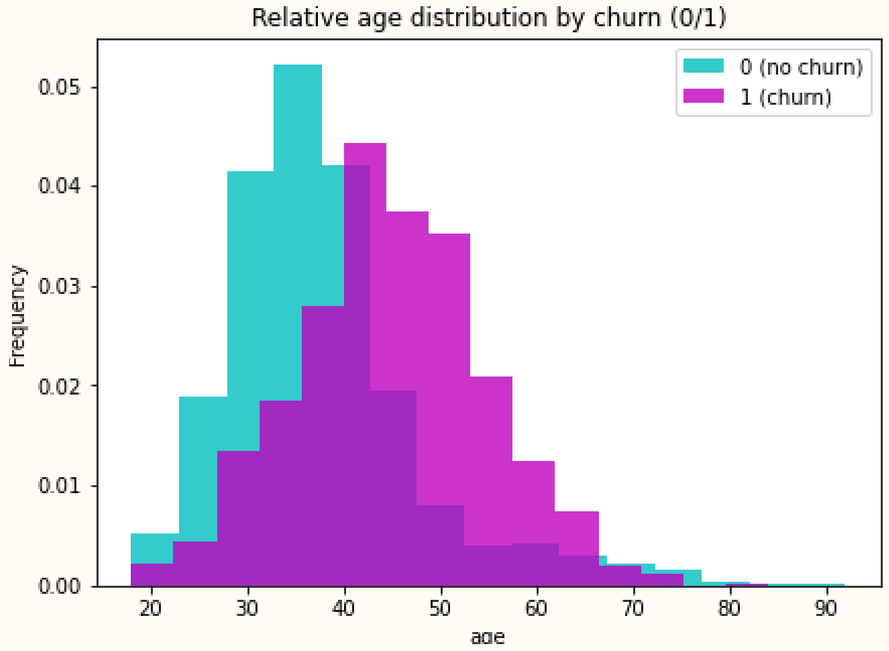


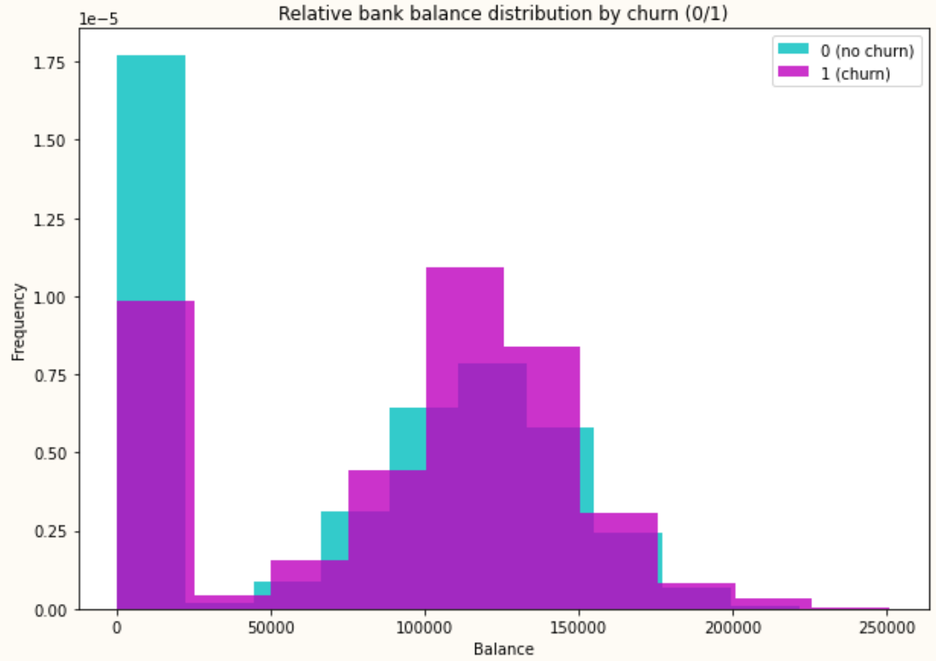
- **Correlations**: In general, there are only a few and very weak correlations shown between variables.
- **Correlated to churn**: age, active_member, balance, products_number.

However, we decided to include all variables in the models we built (except for customers_id).

2. Data Understanding – churn ratio/age/balance

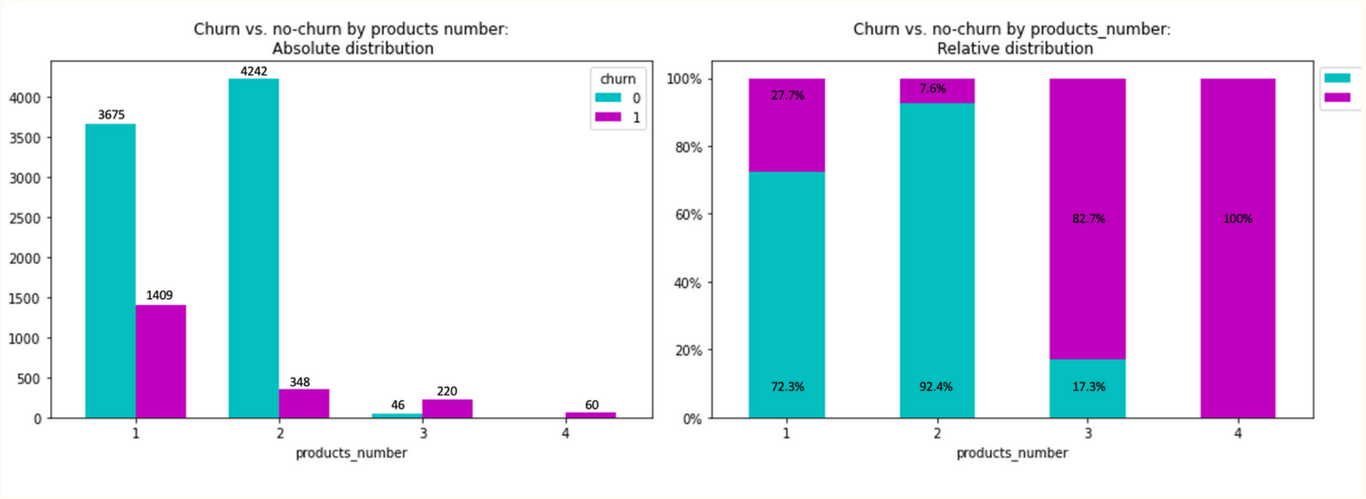




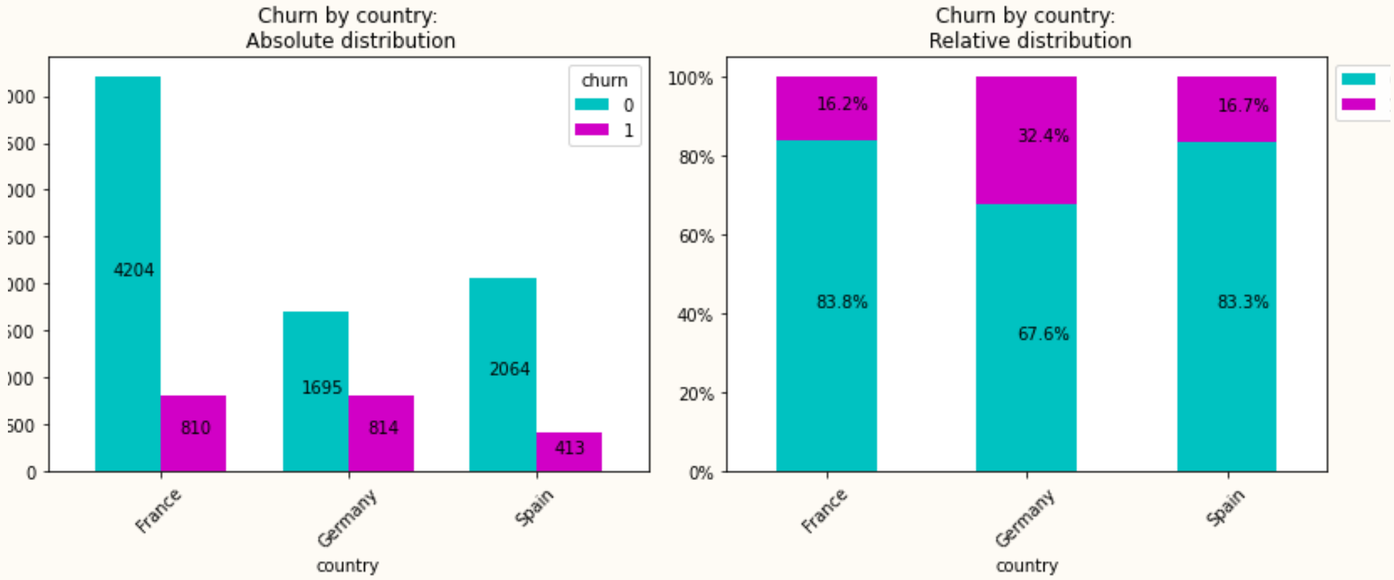


- **Churn rate** is **20.37%.** Out of 10,000 observations, there were 2,037 that churned.
- **Age** has the highest positive correlation with churned customers. **Older customers** were **more likely to churn** than younger ones.
- On average, customers who **churned** tend to have relatively **higher balance** than the ones who stayed.

2. Data Understanding – number of products/country

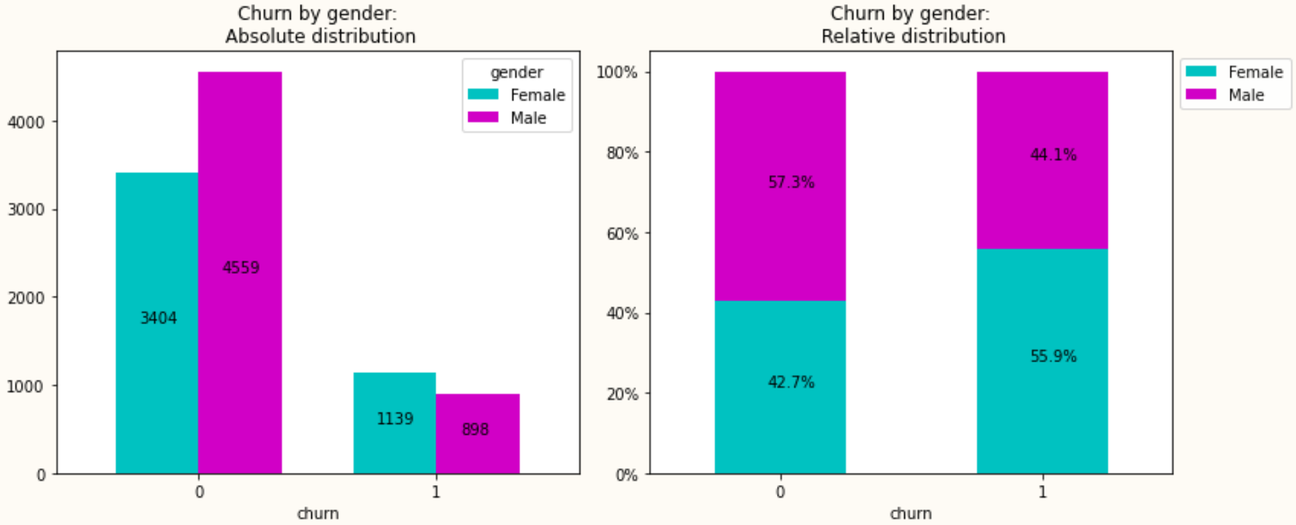


- The churn rate is significantly **higher** when customers start using **more than 2 products** from the bank.
- Noteworthy that the proportion of customers using 3 or 4 products is relatively small.

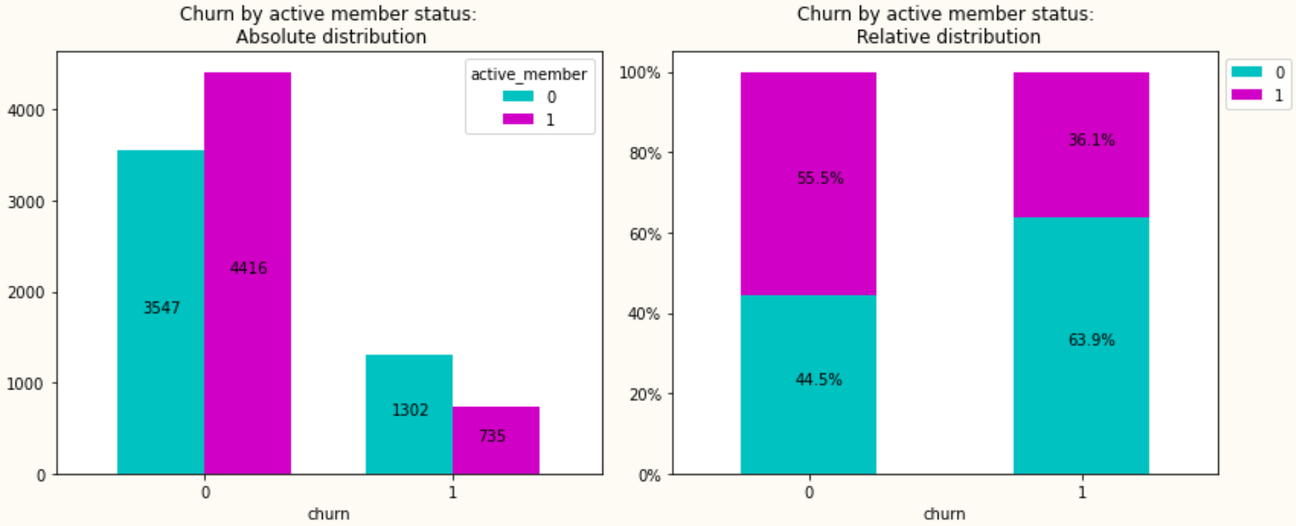


- Three countries: France, Germany, and Spain
- The distribution of samples in each country is unbalanced.
- Customers in **Germany** are **more likely to churn** than those in France or Spain (Germany has the highest absolute churn amount (814) and the highest relative churn rate (32.4%) even though there are more customers in France).

2. Data Understanding – gender/active membership

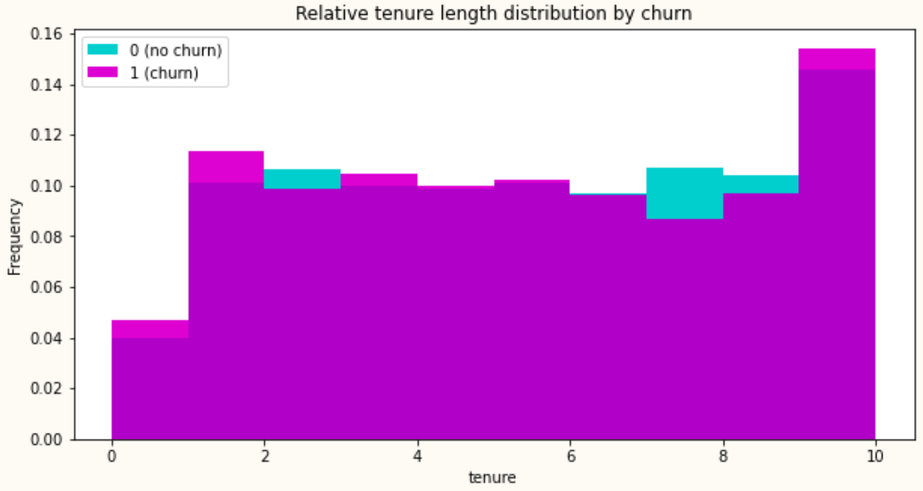


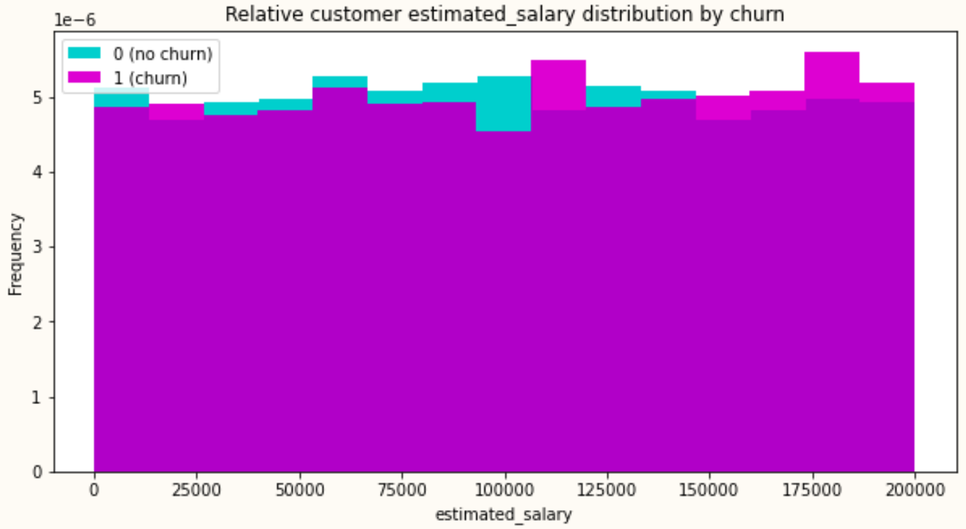
- **Female** customers are **more likely to churn** than male customers**.**

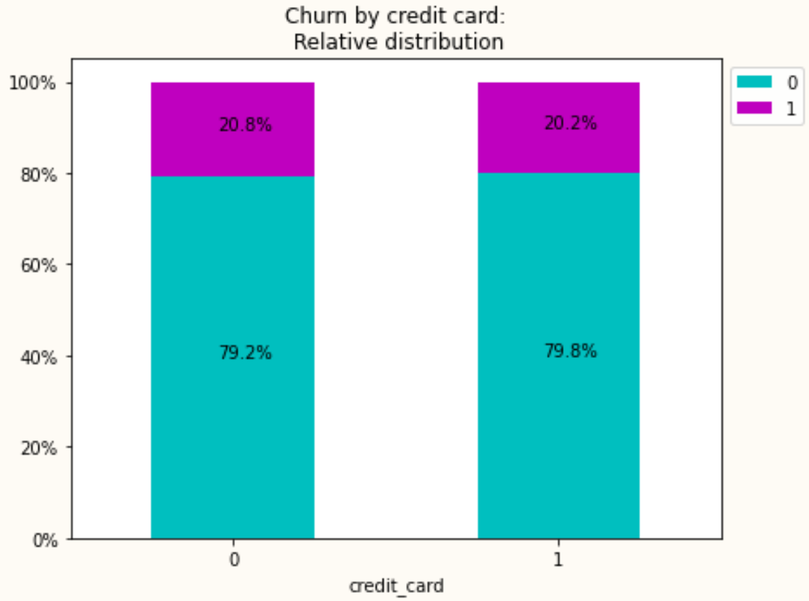


- Overall, **active members** have a **lower churn rate.**

2. Data Understanding – tenure/salary/credit card use







- The churn rate **does not significantly vary** among different **tenure years** or **salaries**.
- There is also **not a remarkable difference** in churn rate, considering whether the customer had a **credit card** or not.

3. Modeling

The dataset is divided into Training Data (***70%***) and Testing Data (***30%***) using **train_test_split** function from **sklearn**.

Due to the **imbalance** in the dataset (20% churn and 80% stay) and we are aiming to predict the **minority** class, we have decided to build 3 prediction models using **both** the **original** and **rebalanced** dataset (using SMOTE method).

- Decision Tree
- Logistic Regression
- Support Vector Machine

The default parameters we used for Decision Tree models are:

- *criterion = gini*
- *max_depth = 3*
- *min_sample_leaf = 3*

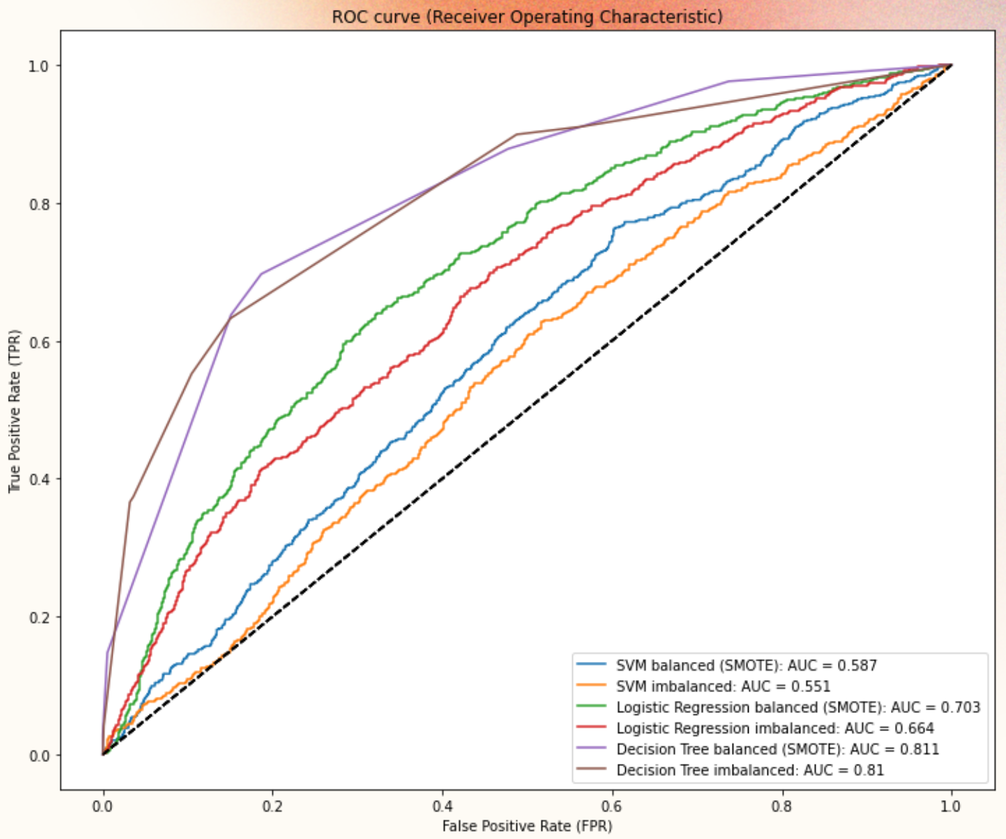
*Visuals of these Decision Tree Models can be found in **Appendix 1**.*

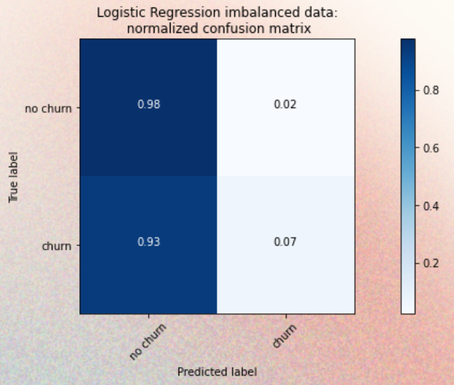
The accuracy scores are listed below, ranked from highest to lowest

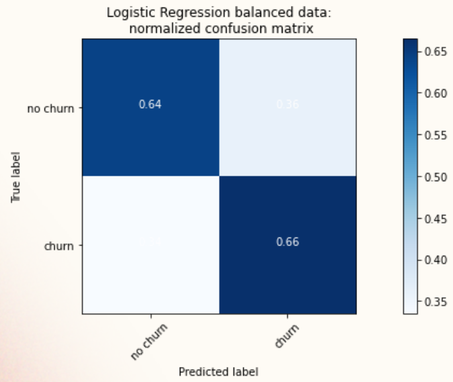
- Decision Tree (unbalanced): 84.30
- Decison Tree (SMOTE): ±79.10
- Logistic Regression (unbalanced): 78.70
- Logistic Regression (SMOTE): ±64.87
- SVM (unbalanced): 79.23
- SVM (SMOTE) ±47.23

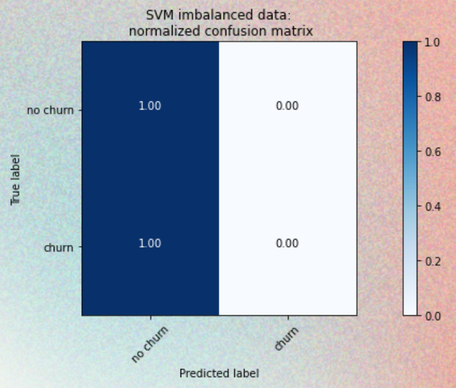
4. Evaluation

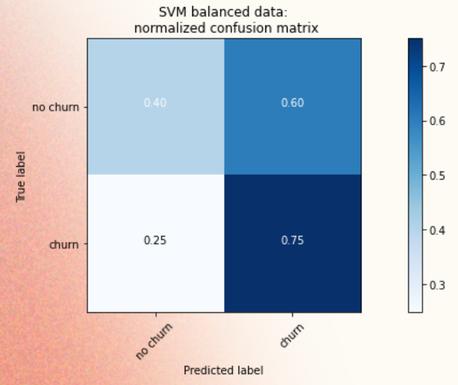
According to the results from ROC curve, AUC score, and confusion matrix, we conclude that **Decision Tree** is the best prediction method for this dataset with result from unbalanced and rebalanced data performing better than other models.

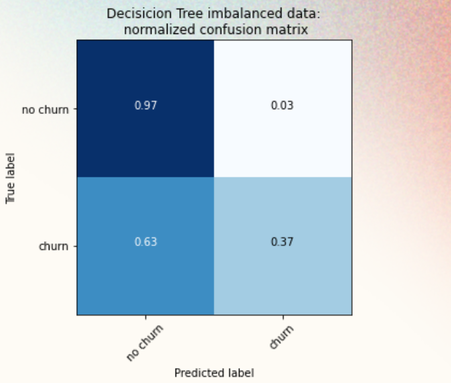


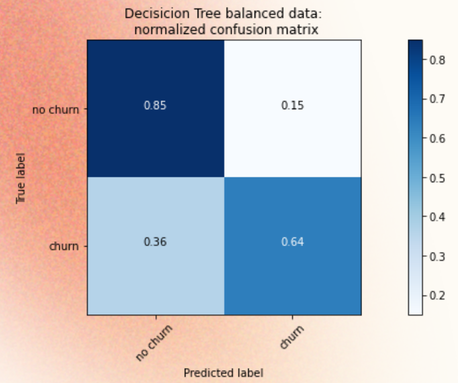




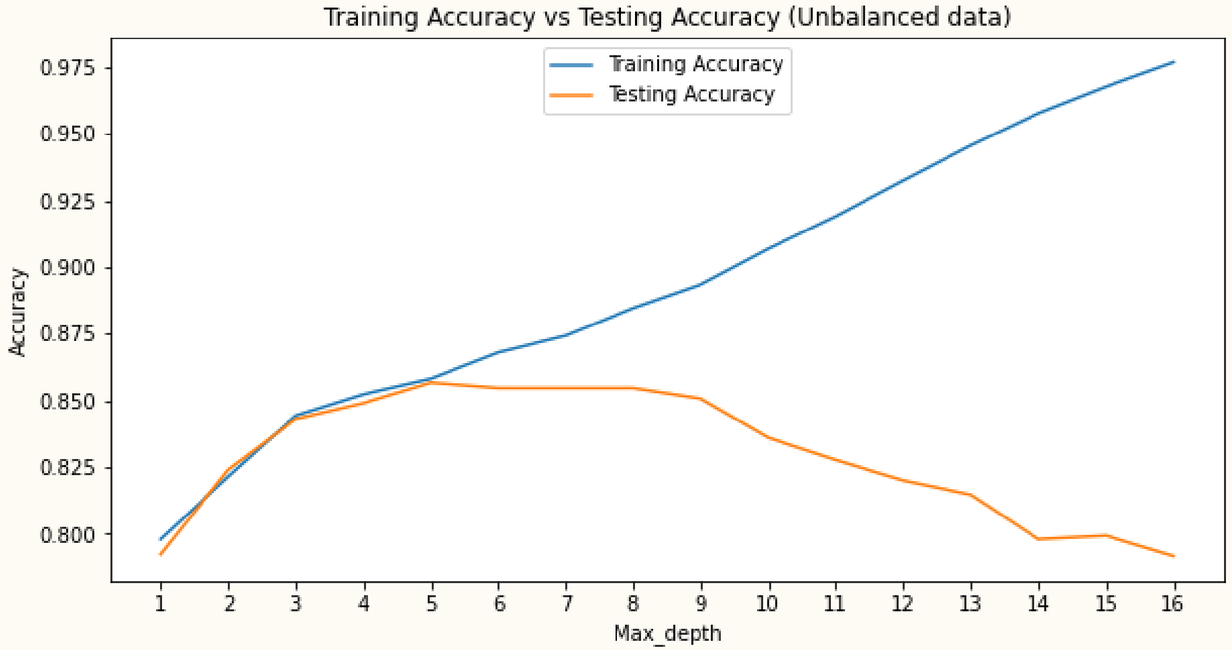


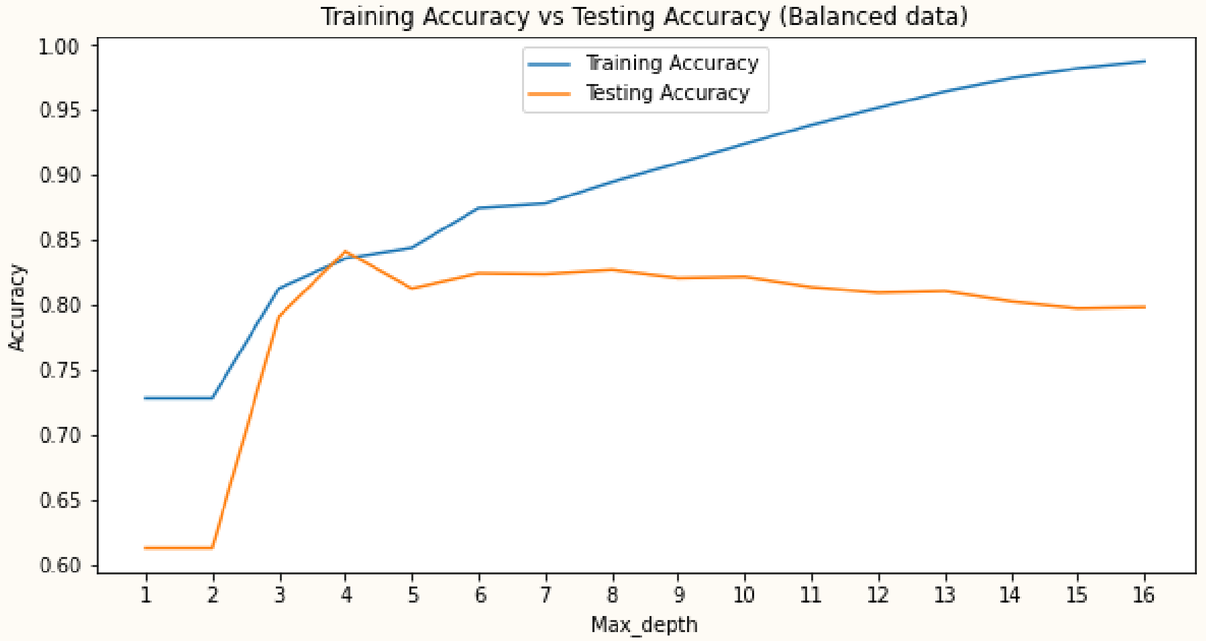






5. Enhancing Decision Tree





- With our chosen Decision Tree model, we decided to execute hyperparamater tuning for **max_depth** values to enhance the model's accuracy.
- According to the charts, we see that for **unbalanced** dataset, **max_depth=5** yields the best accuracy on testing data while for **rebalanced** dataset, **max_depth=4** yields the most accurate predictions on testing data.
- Visuals of the Decision Tree model with new max_depth values can be found in ***Appendix 2***.
- Comparison of ROC curve and AUC score for the Decision Tree models before and after parameters tuning can be found in ***Appendix 3*** showing improvements.

6. Recommendations

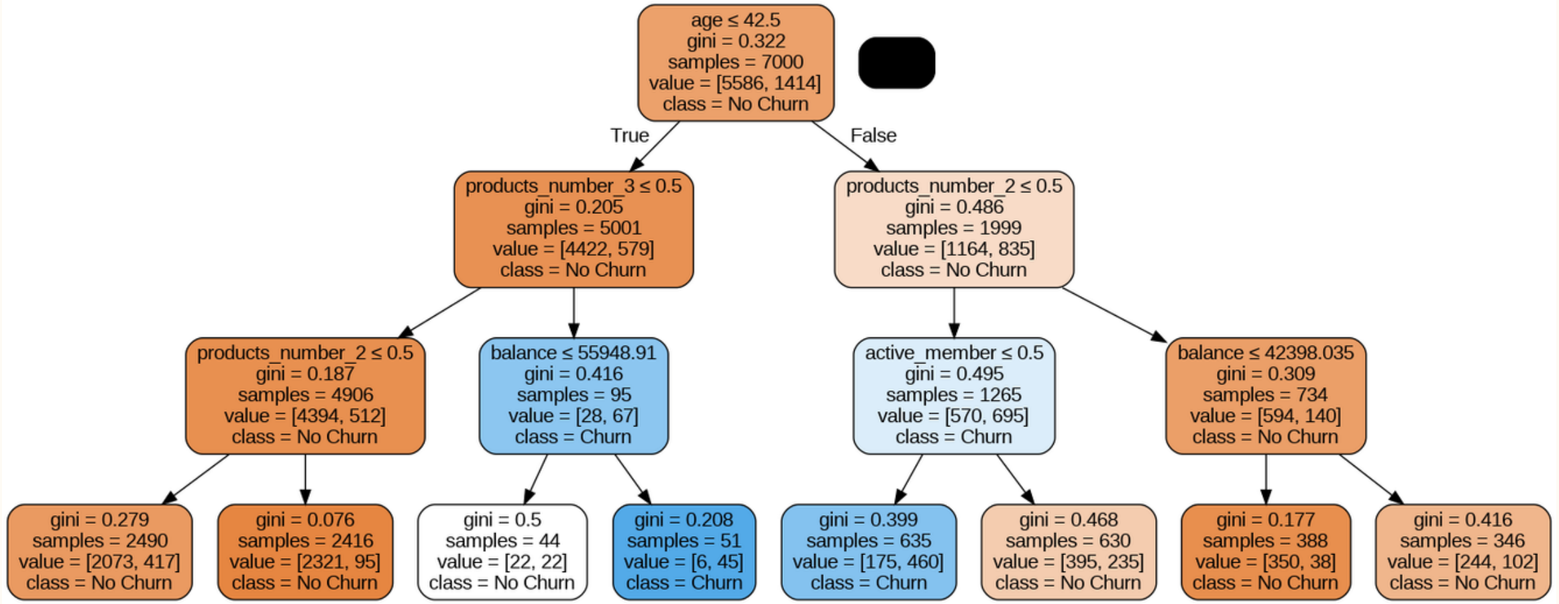
The model that was built can help the bank predict if a customer is more likely to churn - consequently offering a method to more correctly targeted certain groups with a higher risk of churn. For instance, **customers over 42.5** and **inactive customers** are more likely to churn.

Targeted measures can be taken to increase customer experience and reduce the likelihood of churn, for instance:

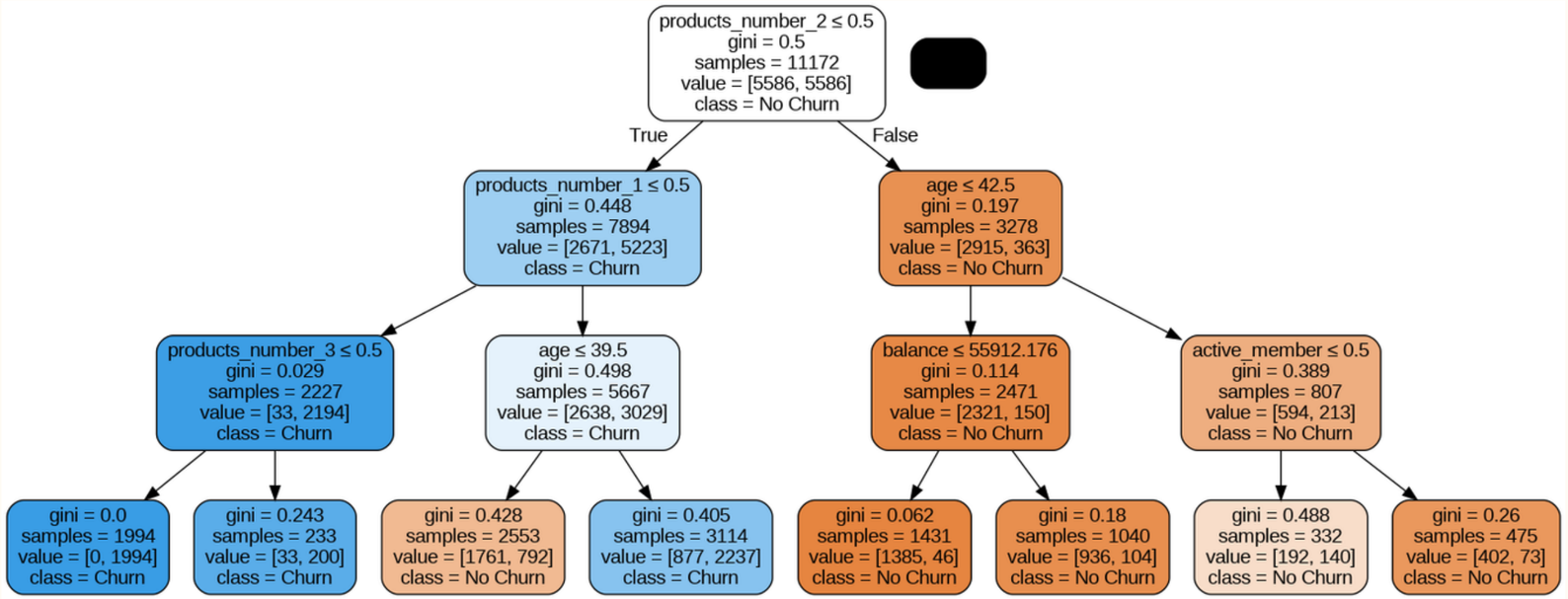
- Take into account churn prediction when defining target customers and implement marketing plans targeting those group with higher churn rates;
- Having tailored rates and fees for different customer groups to increase retention rate;
- Create more incentive programs to make customers become more active users.

For a more accurate prediction model, we suggest the bank further gather and analyze data from more variables that might affect the churn rate to better target the appropriate customer groups and implement more targeted measures, ie. customer experience with in-person services as well as online/mobile services, etc.

Appendix 1



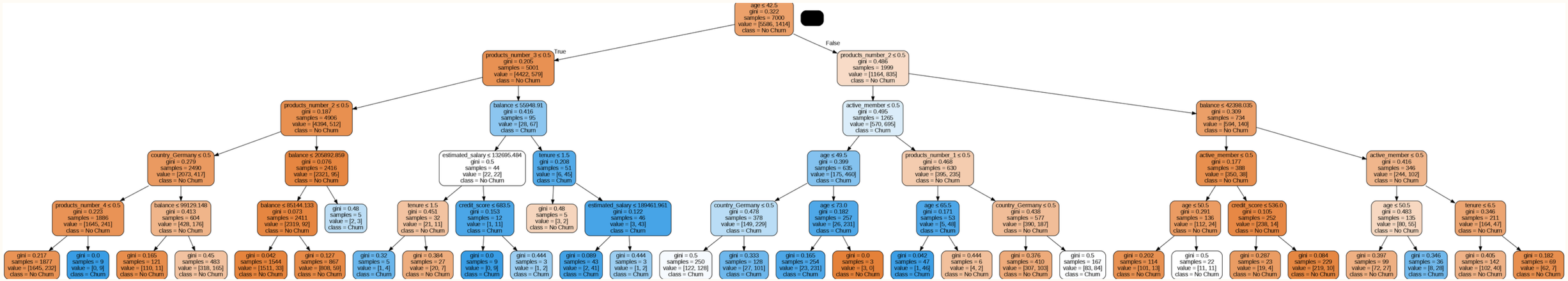
Decision Tree Model from **Imbalanced** data before tuning



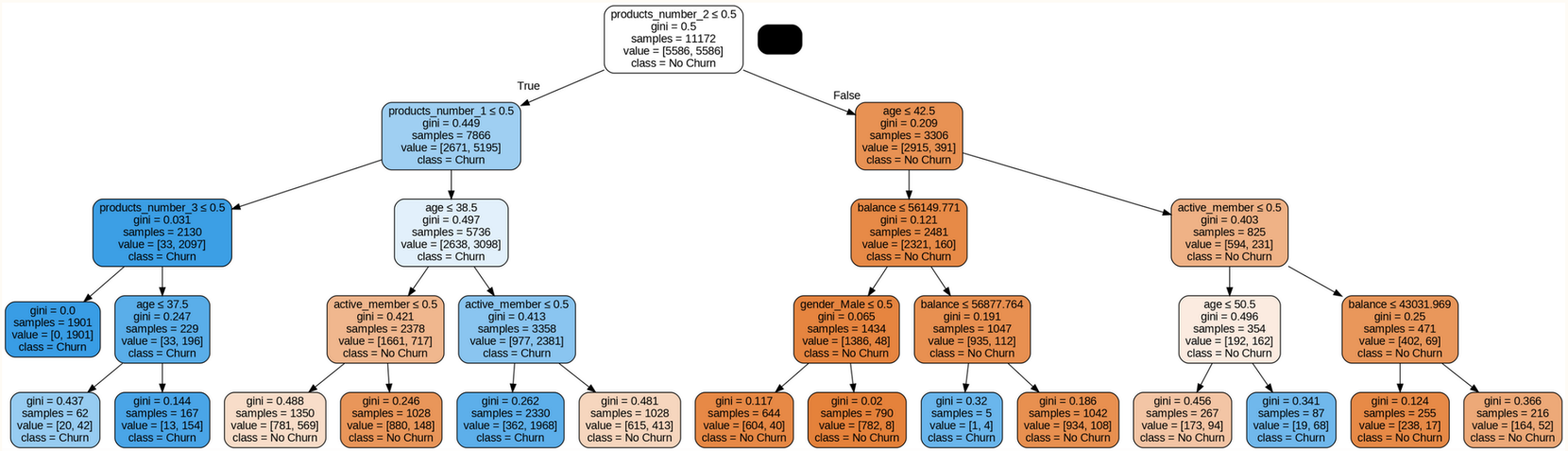
Decision Tree Model from **balanced** data before tuning

Appendix 2

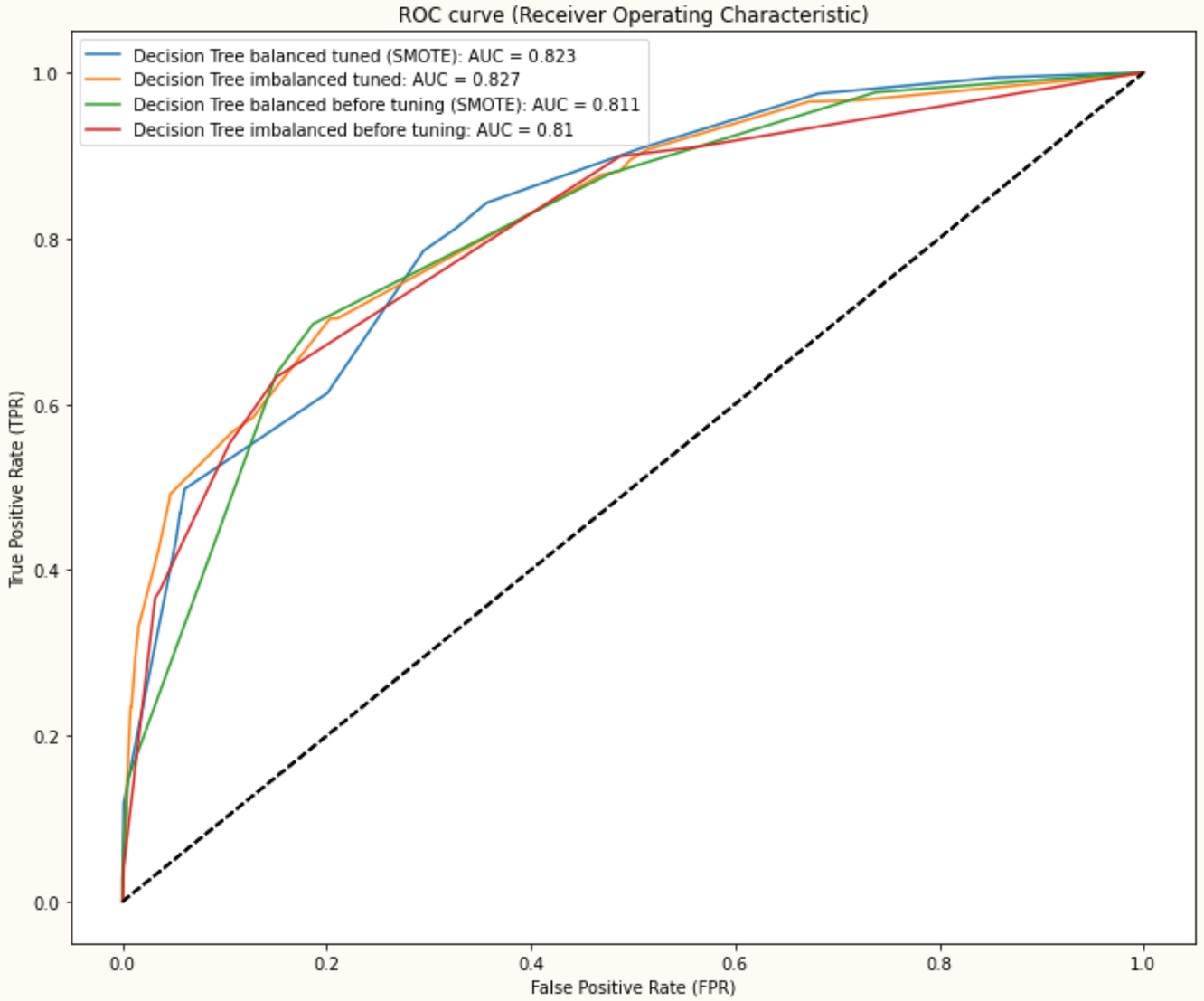
Decision Tree Model from **Imbalanced** data after tuning



Decision Tree Model from **balanced** data after tuning



Appendix 3



Comparison of ROC curve and AUC score of Decision Tree models before and after tuning