# Customer Churn Prediction

# **Saurabh Biswas DSC680 T301** <https://github.com/saurabhbiswas1985/Data-Science/tree/main/Customer%20Churn%20Prediction>

# Any surprises from your domain from these data?

I’m using data from banking domain. My dataset contains a dataset that contains details of banking customer account details and an indicator whether a customer has closed the account or still an active customer. The aim of this project to perform a predictive churn analysis using this banking dataset.

This dataset contains details about 10000 customers out of which 7963 customer has an active account and 2037 customer has closed their account. As we can see, this dataset is not balanced – 79.63% is majority class and 20.37% is minority class.

Working with imbalanced dataset has its won challenge. It requires different metrics to compare the efficiency between different models.

# The dataset is what you thought it was?

I wanted to gather the information regarding bank customers details like age, gender, demographics, credit score, account tenure, balance etc. My aim is to find the impact of these attributes on customer churn. My collected dataset contains 10000 customer details.

Attributes RowNumber, CustomerId, Surname etc. are not relevant for my analysis as these are mostly unique identifier of the customer and doesn’t add any value to churn analysis.

1. CreditScore: ranges between 350 to 850. Shows customer’s creditworthiness.
2. Geography: customers country
3. Gender: Male/Female
4. Age: ranges between 18 to 92 years
5. Tenure: 0 through 10
6. Balance: ranges between 0 to 250000
7. NumOfProducts: count of banking products held by the customer
8. HasCrCard: customer has credit card or not.
9. IsActiveMember: Is customer active.
10. EstimatedSalary: salary of the customer
11. Exited: Customer closed the account or not

I don’t see any missing data in any of these attributes. So overall, this dataset is good to use for carrying out subsequent analysis.

# Have you had to adjust your approach or research questions?

This dataset is imbalanced in nature. The minority class is harder to predict because there are few examples of this class, by definition. This means it is more challenging for a model to learn the characteristics of examples from this class, and to differentiate examples from this class from the majority class.

I’m planning to oversample the minority class to get a balance, before I train any model.

My plan is to use Artificial Neural Network, Random Forest, Gradient boosting, and XGB classifier. But this dataset is not very large. Generally, ANN training requires large datasets. So, in this case, ANN may not provide competitive performance.

# Is your method working?

I used pandas profiling to get an overview of the data. No missing values in any of the attributes.

I have changed values of attributes gender & Geography from string to number but kept the datatype as categorical. This will help me in further model training.

I have dropped RowNumber, CustomerId, Surname columns.

From correlation heatmap, I have identified key attributes that have higher impact on outcome variable ‘Exited’.

I will oversample minority class i.e., customer that closed their accounts to get a more balanced dataset. Then I will split the dataset into test and training dataset and use the training dataset to train my models.

# What challenges are you having?

The research dataset contains only 10000 customer’s data. Training an ANN model requires large datasets, otherwise the performance of this model won’t be good. As I have a small training dataset, I could impact ANN model performance.

I have found that this dataset is imbalanced. Based on my past experience, imbalance dataset prediction model won’t perform very well. My challenge is to build a model that achieves a higher f1-score.