# Customer Churn Prediction

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

# Which Domain?

I’m planning to use data from banking domain. I will use 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.

For any service or product-based organization customer churn is a major issue. When a customer ends a subscription or service and decides to use another service provider, it hits the company’s revenue directly. Customer retention is a key to grow the business. The key to retaining existing customer is to understand the grey areas that customer not happy about and identify high risk group of customers who have high chance to leave. Keeping the high-risk group happy by giving some special perks, can help retaining them.

The aim of this project is to carry out a churn analysis using a banking customer dataset and build a predictive model.

Below are some my source of data and some sources of some similar kind of project done:

Kaggle: <https://www.kaggle.com/shrutimechlearn/churn-modelling?select=Churn_Modelling.csv>

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*(3), 5445-5449.

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547-12553.

Lu, N., Lin, H., Lu, J., & Zhang, G. (2012). A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, *10*(2), 1659-1665.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E.

(2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12,

2825-2830.

Mutanen, T. (2006). Customer churn analysis–a case study. Journal of Product and Brand

Management, 14(1), 4-13.

# Which Data?

##### I will be using bank customer dataset Churn\_Modelling.csv from Kaggle.

Link to the dataset: <https://www.kaggle.com/shrutimechlearn/churn-modelling?select=Churn_Modelling.csv>

This dataset contains details of a bank's customers and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed his account) or he/she continues to be a customer.

The dataset contains 10,000 records with 13 attributes and one target variable.

Demographically the information is about customers from Spain, France, and Germany.

55% of customers are Male, and 45% are Female, with an average age of 38.9 years.

RowNumber

CustomerId

Surname

CreditScore

Geography

Gender

Age

Tenure

Balance

NumOfProducts

HasCrCard

IsActiveMember

EstimatedSalary

Exited

# Research Questions? Benefits? Why analyze these data?

It is natural for companies to focus on getting new customers for growth. But it is equally important to retain their existing customer. Otherwise, it can hit a company’s revenue seriously. According to an internet study 97% of customers churn silently without providing any explanation. It is tough for a company to keep up with customer churn and make a profit at the same time.

Here are some basic questions on this dataset?

* Why a customer closes his account?
* Is there any correlation with age and customer churn?
* Is there any relation with gender and customer churn?
* Does it have any positive effect of holding other products like credit card on keeping customers bank account active?
* How balance affects churn.

# What Method?

I will follow CRISP-DM methodology for this project.

I will first drop unnecessary attributes like RowNumber, CustomerId, Surnameetc. as these won’t help in performing churn analysis. I will analyze the data using pandas profiling and find out whether any attributes have null values or not. Then I will replace those null values with blanks or zeros wherever is applicable.

I will perform with EDA by exploring summary statistics of data. I’m planning to do visualization for

correlation between variables, exited, vs tenure, gender, age, estimated salary, etc.

I’m planning to Artificial Neural Network, Random Forest, Gradient boosting, and XGB classifier and compare their performance.

# Potential Issues?

I don’t see any potential issue at this point of time.

# Concluding Remarks

The aim of this project is to analyze bank customer data and gain insight of the reasons of customer churn. It is important to focus on customers identified by the churn model. Bank and

other financial institutions can greatly benefit from churn analysis. It will help them to retain their existing customer base.