BANK CHURN MODEL

# TEAM NAME: UNIQUE CODERS

# TEAM MEMBERS: SAI RAHUL DODDA

# PINIGANTI KRISHNA VAMSI

# PRATHIPATI SAI NAVEEN

CONTENT PAGE

## PG NO:

# BANK CHURN MODELLING

## INTRODUCTION 2

## OBJECTIVES OF RESEARCH 4

## PROBLEM STATEMENT 6

## INDUSTRY PROFILE 7

# REVIEW OF LITERATURE 9

# METHODOLOGY

## 3.1 EXPLORATORY DATA ANALYSIS 10

# 3.1.1 FIGURES AND TABLES

## 3.2 DATA MODELING USING SUPERVISED ML TECHNIQUES 21

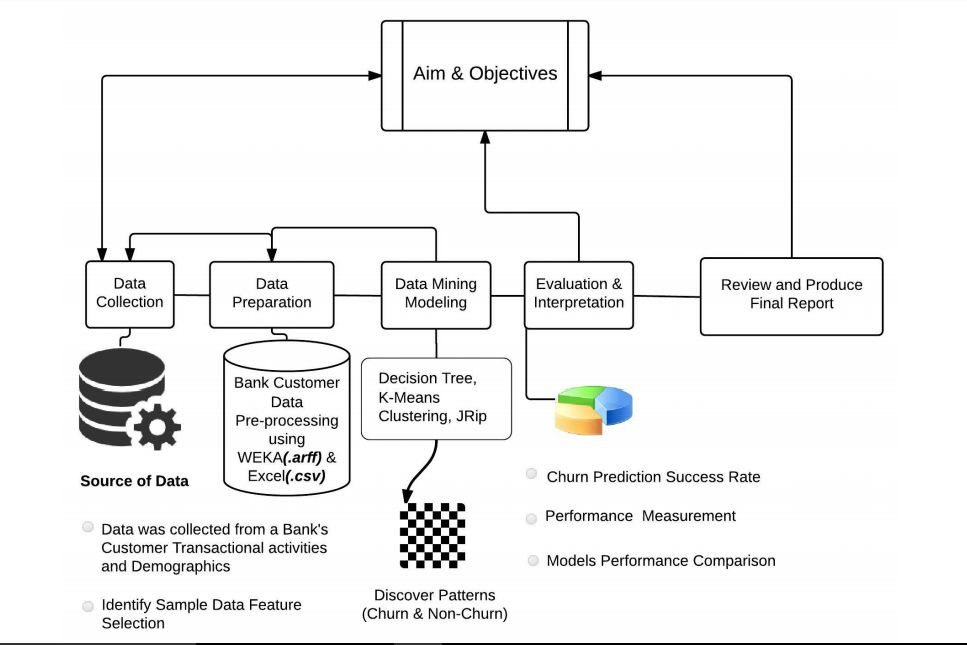
# 4. FINDINGS AND SUGGESTION 24

## 5. CONCLUSION 24

# BANK CHURN MODELLING:

## INTRODUCTION:

With increased availability of data, inexpensive storage and processing power, the amount of raw data stored in banking databases is huge and constantly increasing. However, raw data by itself does not provide much information. Data mining is used to discover patterns and relationships in data in order to improve business decision processes. Its tools can answer business questions that were too time consuming to resolve in the past. We can define it as an interdisciplinary field that brings together techniques from machine learning, pattern recognition, statistics, database systems, data visualization, information theory, knowledge acquisition, artificial intelligence and neural networks. Specific uses of data mining include: Market segmentation, Customer churn, Fraud detection, direct marketing, Interactive marketing, Market basket analysis, Trend analysis, Credit analysis, Predicting payment default, etc.



In this paper, we will focus on Customer churn. Techniques that are most commonly used to predict customer churn are: neural networks, support vector machines and logistic regression models. We want to make a model from stored customer data to predict churn and to prevent the customer’s turnover. Data mining research literature suggests that machine learning techniques, such as neural networks should be used for non-parametric datasets, because they often outperform traditional statistical techniques such as linear and quadratic discriminant analysis approaches.

In the era of globalization and intents competition in banking industry, banks are forced to fight more creatively and proactively to gain and maintain their clients. Questions data mining can answer are:

• What transactions does a customer do before shifting to a competitor bank? ,

• Which bank products are often availed of together by which groups of customers? ,

• What patterns in credit transactions show increased risk of fraud? ,

• What is the profile of a high-risk borrower? , And

• What services and benefits would current customers likely desire?

Banks have realized that customer relations are a very important factor for their success. The challenge banks face is how to retain most profitable customers. Literature suggests that a small change in the retention rate can result in significant impact on business. Huge amount of costumer and transaction data are maintained by banks, but because of size of these databases makes it impossible for the banks to analyse and to retrieve useful information for the decision makers. Data mining is a powerful tool that can find patterns and relationships within a data. Using data mining technique, it is possible to build a successful predictive model which transforms data into meaningful information.

This paper proposes a machine learning based approach to predict customer churn in bank. Real-world data from one of the small Croatian banks was used for creating a model for Customer churn. The main hypothesis was that clients who use more bank services (products) are more loyal, and bank should focus on those clients who use less than three products, and offer them products according to their needs

­­­­­­

## 1.2 OBJECTIVES OF RESEARCH:

## The need for customer churn prediction:

Our case data in this paper was provided by a company operating in a retail banking sector. In personal retail banking a company must operate on a long term customer strategy, young customers are recognized as being unprofitable in the early stage in lifecycle but will become profitable later on. So as the customer relationships last, maybe decades, the company must address the value of a potential loss of a customer. The customer lifetime value analysis will help to face this challenge.

## Customer lifetime value concept:

The customer lifetime value is usually defined as the total net income from the customer over his lifetime. This type of customer analysis is done under several terms: customer value, customer lifetime value, customer equity, and customer profitability. The underlying idea in LTV concept is simple and measuring the lifetime value is easy after the customer relationship is over. The challenge in this concept is to define and measure the customer lifetime value during, or even before, the active stage of customer relationship.

A conceptual LTV model is defined as follows:

LTV is the total value of direct contributions and indirect contributions to overhead and profit of an individual customer during the entire customer life cycle that is from start of the relationship until its projected ending.

Most LTV models stem from the basic equation, although there are also many other LTV models having various application areas. The components of the basic LTV model are [3]:

- The customer net present value over time (revenue and cost).

- Retention rate or length of service (LOS).

- Discount factor.

Each component can be measured or estimated separately and then combined for the LTV model.

The benefits of better understanding the customer lifetime value are numerous. The company can measure the present and the future income from the customers. The company can also foster customer retention and loyalty which will lead to higher customer profitability. The LTV analysis can also help the company on their customization of products and services. This understanding of the customer value helps the company to focus on revenue productive customers and yield the customer segment with potential negative impacts to the revenue. And last, the customer lifetime value is not a fixed value it can be influenced by marketing efforts.

## Customer churn

The focus on customer churn is to determinate the customers who are at risk of leaving and if possible on the analysis whether those customers are worth retaining. The churn analysis is highly dependent on the definition of the customer churn. The business sector and customer relationship affects the outcome how churning customers are detected. Example in credit card business customers can easily start using another credit card, so the only indicator for the previous card company is declining transactions. On the other hand for example in Finnish wireless telecom industry a customer can switch one carrier to another and keep the same phone number. In this case the previous carrier will get the signal right at the churning moment. The customer churn is closely related to the customer retention rate and loyalty. Their LTV suggest that churn rate of a customer has strong impact to the LTV value because it affects the length of service and the future revenue. Hwang et al. also defines the customer loyalty as the index that customers would like to stay with the company. Churn describes the number or percentage of regular customers who abandon relationship with service provider.

## 1.3 PROBLEM STATEMENT:

Churn prediction is one of the common applications of the classification in the business settings. The word “churn” means to stop consuming products of a specific company and use fungible product of another company because of its better quality or service or less price. There are lots of studies which show that acquiring a new customer for a company is five or six times more expensive than retaining an existing one. Accordingly, nowadays most of the financial institutions are concerned with customer retention studies to prevent losing their market share and maximize their gained profit from existing customers. The primary objective of customer retention is to maximize the potential profit which can come from existing customers. In most of the churn prediction studies, the objective of classification Profit-based classification in customer churn prediction: a case study in banking industry 2 is to minimize the prediction error and accordingly maximize the accuracy of the prediction. This approach is definitely an optimal approach when the objective is to correctly classify the customers as much as possible, however, it may reach suboptimal solution when the objective is to maximize the profit of churn prediction for the company. In our case, the bank has information about customers’ lifetime value for the next period (one year) which can be used as a profit metric to show the importance of each of the customers. In this study, we have two objective:

a. Developing a profit-based classification algorithm which classifies churners and non-churners such that it maximizes the total potential profit of the bank by giving more weight to detection of profitable churner customers.

b. Finding appropriate individual incentive offer value for each of the churner customers instead of giving fixed offers to all of them to ensure that more profitable customers are getting more valuable offers than other churners and accordingly minimize their corresponding churn (leaving) probability.

## INDUSTRY PROFILE:

A bank is a commercial or state institution that provides financial services, including issuing money in form of coins, banknotes or debit cards, receiving deposits of money, lending money and processing transactions.

A commercial bank accepts deposits from customers and in turn makes loans based on those deposits.

Some banks (called Banks of issue) issue banknotes as legal tender. Many banks offer ancillary financial services to make additional profit.

For example: selling insurance products, investment products or stock broking.

Most banks also rent safe deposit boxes in their vault.

Currently in most jurisdictions, commercial banks are regulated and require permission to operate.

Operational authority is granted by bank regulatory authorities and provide rights to conduct the most fundamental banking services such as accepting deposits and making loans.

A commercial bank is usually defined as an institution that both accepts deposits and makes loans; there are also financial institutions that provide selected banking services without meeting the legal definition of a bank (see banking institutions).

Banks have a long history, and have influenced economies and politics for centuries.

In history, the primary purpose of a bank was to provide liquidity to trading companies.

Banks advanced funds to allow businesses to purchase inventory, and collected those funds back with interest when the goods were sold.

For centuries, the banking industry only dealt with businesses, not consumers.

Commercial lending today is a very intense activity, with banks carefully analysing the financial condition of its business clients to determine the level of risk in each loan transaction.

Banking services have expanded to include services directed at individuals and risks in these much smaller transactions are pooled.

A bank generates a profit from the differential between what level of interest it pays for deposits and other sources of funds, and what level of interest it charges in its lending activities.

This difference is referred to as the spread between the cost of funds and the loan interest rate.

Historically, profitability from lending activities has been cyclic and dependent on the needs and strengths of loan customers.

In recent history, investors have demanded a more stable revenue stream and banks have therefore placed more emphasis on transaction fees, primarily loan fees but also including service charges on array of deposit activities and ancillary services (international banking, foreign exchange, insurance, investments, wire transfers, etc.).

However, lending activities still provide the bulk of a commercial bank's income.

The name bank derives from the Italian word banco, desk, used during the Renaissance by Florentines bankers, who used to make their transactions above a desk covered by a green tablecloth

# review of literature:

Previous studies have identified the benefits that customer retention delivers to an organisation. For example, the longer a customer stays with an organisation the more utility the customer generates. This is an outcome of a number of factors relating to the time the customer spends with the organisation. These includes the higher initial costs of introducing and attracting a new customer, increases in both the value and number of purchases, the customer’s better understanding of the organisation, and positive word -of – mouth promotion.

Apart from the benefits that the longevity of the customers brings, research findings also suggest that the costs of customer retention activities are less than the costs of acquiring new customers. From the previous studies the financial implications of attracting new customers may be five times as costly as keeping existing customers. Banks loses satisfied customers who have moved, retired, or no longer need certain services. As a consequence, retaining customers becomes a priority.

Clearly, there are compelling arguments for bank management to carefully consider the factors that might increase customer retention rates. However, there has been little empirical research that investigates the constructs leading to customer retention. Examples of constructs are competitive advantage, customer satisfaction, switching barriers, corporate image, and bank services characteristics. There have been few, if any, attempts to link them to customer retention. This is curious, for if retention criteria are not well manage, customers might still leave their banks, no matter how hard bankers try to retain them.

# methodology:

## 3.1 EXPLORATORY DATA ANALYSIS:

### In statistics, exploratory data analysis (EDA) is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. Exploratory data analysis was promoted by John Turkey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA. Our data set contains the attributes RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary and Exited.

### FIGURES AND TABLES:

#### Checking of null values:

The number of null values is:

RowNumber 0

CustomerId 0

Surname 0

CreditScore 0

Geography 0

Gender 0

Age 0

Tenure 0

Balance 0

NumOfProducts 0

HasCrCard 0

IsActiveMember 0

EstimatedSalary 0

Exited 0

dtype: int64

#### Data types of each attribute:

RowNumber int64

CustomerId int64

Surname object

CreditScore int64

Geography object

Gender object

Age int64

Tenure int64

Balance int64

NumOfProducts int64

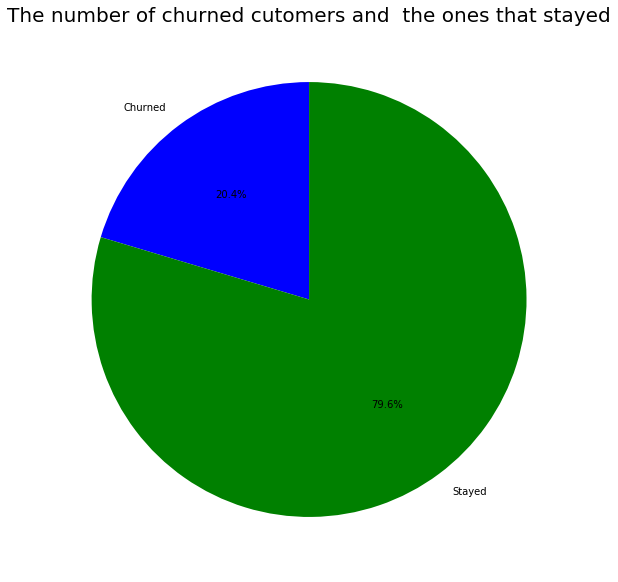
HasCrCard int64

IsActiveMember int64

EstimatedSalary int64

Exited int64

dtype: object

1. Churned vs stayed customers:
2. Displaying the first five rows of objects using head() :

Surname Geography Gender

0 Hargrave France Female

1 Hill Spain Female

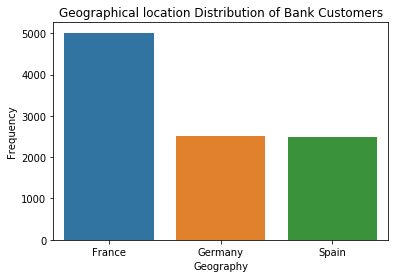
2 Onio France Female

3 Boni France Female

4 Mitchell Spain Female

The number of null values is: 0

#### Comparing the geographical locations of the customers



#### Comparison of gender of the customers:

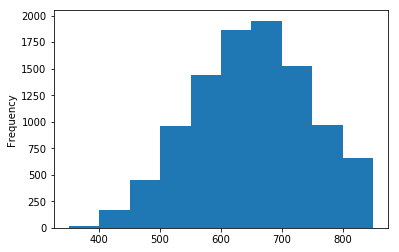


1. Graph for displaying rownumber:

#### Graph representing the customer id:



#### Graph representing credit score:

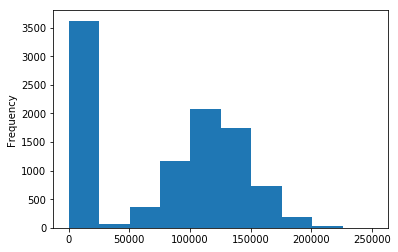


#### Graph representing the age of the customers:

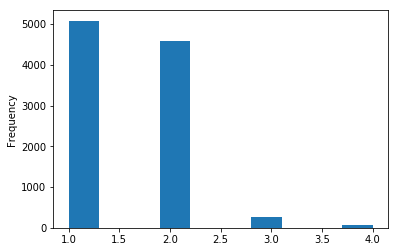
#### Graph representing the tenure of the customers:



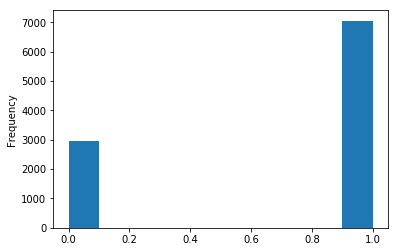
#### Graph representing the balance of the customers:



#### Graph representing the number of products used by the customers:



#### Graph representing the number of customers having credit card:



#### Graph representing the number of active customers:



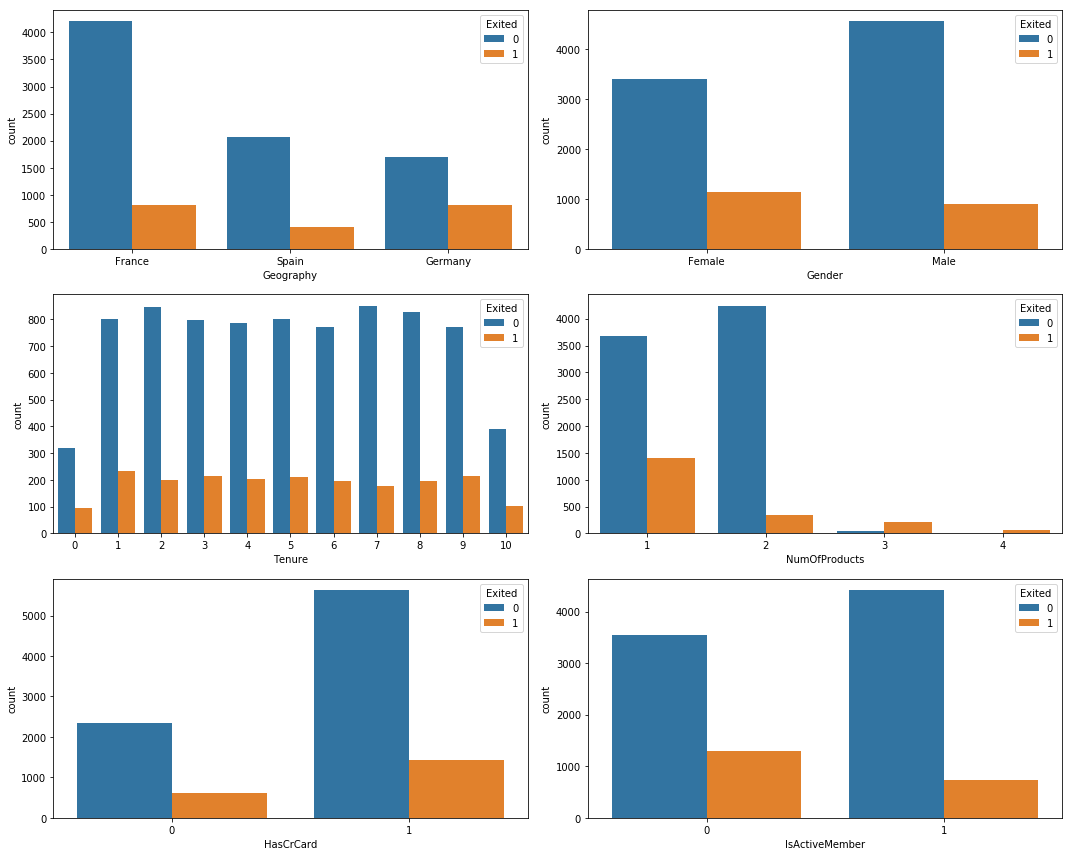
#### Graph representing the estimated salary of the customers:



#### Graph representing the number of exited customers:



#### Graph comparing exited customers with respect to their attributes:



#### Table representing the change of strings into binary format:

RowNumber CustomerId Surname CreditScore Geography Gender Age \

0 1 15634602 Hargrave 619 France 0 42

1 2 15647311 Hill 608 Spain 0 41

2 3 15619304 Onio 502 France 0 42

3 4 15701354 Boni 699 France 0 39

4 5 15737888 Mitchell 850 Spain 0 43

Tenure Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary \

0 2 0 1 1 1 101348

1 1 83807 1 0 1 112542

2 8 159660 3 1 0 113931

3 1 0 2 0 0 93826

4 2 125510 1 1 1 79084

Geography\_France Geography\_Germany Geography\_Spain Exited

0 1 0 0 1

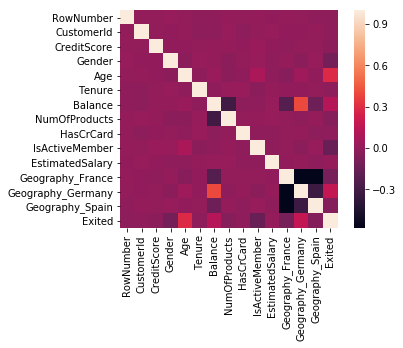
1 0 0 1 0

2 1 0 0 1

3 1 0 0 0

1. 0 0 1 0

#### Correlation heat map:



## 3.2) DATA MODELING USING SUPERVISED ML TECHNIQUES:

**Random forests** or **random decision forests** are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [over fitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

The first algorithm for random decision forests was created by [Tin Kam Ho](https://en.wikipedia.org/wiki/Tin_Kam_Ho) using the [random subspace method](https://en.wikipedia.org/wiki/Random_subspace_method), which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by [Leo Breiman](https://en.wikipedia.org/wiki/Leo_Breiman) and Adele Cutler, who registered"Random Forests" as a [trademark](https://en.wikipedia.org/wiki/Trademark) (as of 2019, owned by [Minitab, Inc.](https://en.wikipedia.org/wiki/Minitab)). The extension combines Breiman's "[bagging](https://en.wikipedia.org/wiki/Bootstrap_aggregating)" idea and random selection of features, introduced first by Ho and later independently by Amit and [Geman](https://en.wikipedia.org/wiki/Donald_Geman" \o "Donald Geman) in order to construct a collection of decision trees with controlled variance

### Preliminaries: decision tree learning

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say [Hastie](https://en.wikipedia.org/wiki/Trevor_Hastie) *et al.*, "because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of features, and produces inspectable models. However, they are seldom accurate".

In particular, trees that are grown very deep tend to learn highly irregular patterns: they [over fit](https://en.wikipedia.org/wiki/Overfitting) their training sets, i.e. have [low bias, but very high variance](https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff). Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

### Bagging

The training algorithm for random forests applies the general technique of [bootstrap aggregating](https://en.wikipedia.org/wiki/Bootstrap_aggregating), or bagging, to tree learners. Given a training set *X* = *x1*, ..., *xn* with responses *Y* = *y1*, ..., *yn*, bagging repeatedly (*B* times) selects a [random sample with replacement](https://en.wikipedia.org/wiki/Sampling_(statistics)#Replacement_of_selected_units) of the training set and fits trees to these samples:

For *b* = 1, ..., *B*:

1. Sample, with replacement, *n* training examples from *X*, *Y*; call these *Xb*, *Yb*.
2. Train a classification or regression tree *fb* on *Xb*, *Yb*.

After training, predictions for unseen samples *x'* can be made by averaging the predictions from all the individual regression trees on *x'*:

{\displaystyle {\hat {f}}={\frac {1}{B}}\sum \_{b=1}^{B}f\_{b}(x')}

Or by taking the majority vote in the case of classification trees.

This bootstrapping procedure leads to better model performance because it decreases the [variance](https://en.wikipedia.org/wiki/Bias%E2%80%93variance_dilemma) of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets.

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on *x'*:

{\displaystyle \sigma ={\sqrt {\frac {\sum \_{b=1}^{B}(f\_{b}(x')-{\hat {f}})^{2}}{B-1}}}.}

The number of samples/trees, *B*, is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal number of trees *B* can be found using [cross-validation](https://en.wikipedia.org/wiki/Cross-validation_(statistics)), or by observing the [*out-of-bag error*](https://en.wikipedia.org/wiki/Out-of-bag_error): the mean prediction error on each training sample *xᵢ*, using only the trees that did not have *xᵢ* in their bootstrap sample. The training and test error tend to level off after some number of trees have been fit.

### From bagging to random forests

The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a [random subset of the features](https://en.wikipedia.org/wiki/Random_subspace_method). This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) are very strong predictors for the response variable (target output), these features will be selected in many of the *B* trees, causing them to become correlated. An analysis of how bagging and random subspace projection contribute to accuracy gains under different conditions is given by Ho.

Typically, for a classification problem with *p* features, √*p* (rounded down) features are used in each split. For regression problems the inventors recommend *p/3* (rounded down) with a minimum node size of 5 as the default.. In practice the best values for these parameters will depend on the problem, and they should be treated as tuning parameters.

# 4) FINDINGS AND SUGGESTION:

In this predictive model, we found that credit score, tenure, balance, estimated salary and has a credit card or not are the important independent variables and has a huge impact on the Target variable. If all values of these variables are low then the customer is more likely to churn. If one or two independent variables are having a low values and remaining are having high values then there is a chance that the customer may churn or stay, it depends on the previous fed data set and also the machine chooses the best by analysing the given input data.

# 5) conclusion:

The Final outcome of this task is to find whether the customer would like to churn or exit his bank account and we have done that perfectly by Data pre-processing, visualising the data and by applying different kind of classification algorithms we found that Random forest classifier is the one which gave us the best accuracy of 86%.  Having an accuracy of 86% means that our model is good enough to predict the customer churn even if we give the new data as input.