A Project Report On

**“CUSTOMER CHURN ANALYSIS**

**USING**

**MACHINE LEARNING”**

SUBMITTED IN THE FULFILLMENT OF THE REQUIREMENT

FOR THE AWARD OF THE DEGREE

**Bachelor Of Technology**

**IN**

**INFORMATION TECHNOLOGY**

SUBMITTED BY

**USHNIS PANJA (12500216008)**

**SUGATA KUNDU (12500216019)**

UNDER THE ESTEEMED GUIDANCE OF

**MR. SOUMYAKANTI BHATTACHARYA**

**Asst. Professor**

**Dept. Of CSE & IT**



DEPARTMENT OF INFORMATION TECHNOLOGY

Bengal College Of Engineering & Technology

Durgapur, West Bengal

**CERTIFICATE OF APPROVAL**

The project entitled **“CUSTOMER CHURN ANALYSIS USING MACHINE LEARNING”** submitted by **Ushnis Panja (12500216008)** and **Sugata Kundu (12500216019)** under guidance of **Mr. Soumyakanti Bhattacharya**, is hereby approved as a credible study of engineering study to warrant its acceptance as a pre-requisite to obtain the degree for which it has been submitted. It is understood that by this approval the undersigned don’t necessarily endorse or approve any statement made, opinion or conclusion drawn therein but approve the project only for the purpose for which it is submitted.

Mr. Soumyakanti Bhattacharya

Project Mentor

Asst. Professor, Department Of CSE & IT

**UNDERTAKING**

We, **Ushnis Panja (12500216008**) and **Sugata Kundu (12500216019)** of B.Tech,7th Semester from Information Technology(IT) Department, hereby declare that our project report entitled **“CUSTOMER CHURN ANALYSIS USING MACHINE LEARNING”** is our own contribution. The work or ideas of other people which are utilized in building this project has been properly acknowledged and mentioned in the bibliography. We undertake total responsibility if any traces of plagiarism is found at any later stage.

Ushnis Panja (12500216008)

Sugata Kundu (12500216019)

**ACKNOWLEDGEMENT**

We wish to express our deep sense of gratitude and indebtedness to Mr. Soumyakanti Bhattacharya Sir, faculty of CSE & IT, for the vision and foresight which inspired us to conceive this project.

The architecture of success stands on a strong foundation made up of strenuous hard work, determination, presence of mind and above all, advice from the learned and experienced people. This word of acknowledgement is to express our appreciates to all those luminaries and invisible hands without whose support the completion of this dissertation would not have been materialized.

Finally, we would like to thank our director sir Prof. Dr. A.C. Ganguly, Principal Sir Prof Dr. R. C. Soni and HOD Sir Mr. Biplab Palye Sir for their encouragement and providing facilities needed for the completion of this project.

Date: 30th April 2020

Place: Durgapur

**INDEX**

1. Abstract……………………………………………………………………………………..6

2. Introduction……………………………………………………………………………….7-8

3. Literature Review

3.1The customer lifetime value concept…………………………………………………..9

3.2Customer Churn…………………………………………………………………...10-12

4. Traditional Methods of Solving Customer Churn……………………………………..13-14

5. Proposed Method

5.1Binary Classifier…………………………………………………………………..15

5.2Artificial Neural Network …………………………………………………………16

5.2.1 Perceptrons and Multi-Layer Perceptrons……………….......................17

5.2.2 Learning…………………………………………………………...17 - 19

5.3Confusion Matrix and AUC-ROC Curve……………………………………..20 - 22

6. Results and Discussions

6.1 Case Data…………………………………………………………………….23-24

6.2 Case Focus………………………………………………………………………25

6.3 Predictive Performance and Model Evaluation

6.3.1 Confusion Matrix…………………………………………………25 – 26

6.3.2 AUC-ROC Curve…………………………………………………26 - 28

7. Future Scope and Conclusion…………………………………………………………….29

8. References………………………………………………………………………………...30

**1. ABSTRACT**

Customer value analysis is critical for a good marketing and a customer relationship management strategy. An important component of this strategy is the customer retention rate. Customer retention rate has a strong impact on the customer lifetime value, and understanding the true value of a possible customer churn will help the company in its customer relationship management. Conventional statistical methods are very successful in predicting a customer churn. The goal of this study is to apply classification techniques to predict a customer churn and analyse the churning and no-churning customers by using data from a personal retail banking company.

This report outlines the impact of customer churn, provides the research on customer churn, gives a detailed view about our model regarding prediction and customer churn its scopes and limitations, also discusses whether this type of model is relevant for future use.

**2. INTRODUCTION**

The subject of customer retention, loyalty, and churn is receiving attention in many industries. This is important in the customer lifetime value context. A company will have a sense of how much is really being lost because of the customer churn and the scale of the efforts that would be appropriate for retention campaign. The mass marketing approach cannot succeed in the diversity of consumer business today. Customer value analysis along with customer churn predictions will help marketing programs target more specific groups of customers.

The subject of customer retention, loyalty, and churn is receiving attention in many industries. This is important in the customer lifetime value context. A company will have a sense of how much is really being lost because of the customer churn and the scale of the efforts that would be appropriate for retention campaign. The mass marketing approach cannot succeed in the diversity of consumer business today. Customer value analysis along with customer churn predictions will help marketing programs target more specific groups of customers.

Personal retail banking sector is characterized by customers who stays with a company very long time. Customers usually give their financial business to one company and they won’t switch the provider of their financial help very often. In the company’s perspective this produces a stable environment for the customer relationship management. Although the continuous relationships with the customers the potential loss of revenue because of customer churn in this case can be huge.

**This report will present a customer churn analysis in personal retail banking sector. The goal of this report is twofold. First the churning customers analysed in banking context. The second objective is a forecast of churning customers based on a classification model backboned by artificial neural network.**

**After the introduction this paper has 6 sections. The background for customer lifetime value concept and a literature review about the customer churn included in the section 3. The methods used in this study and the traditional methods are presented in section 4 and section 5 respectively. A closer look at the case data and the focus of this case study along with the results of the churn prediction are presented in the section 6. And conclusions of this study are left in section 7. All the links of the references have been given in section 8.**

**3. THE NEED FOR CUSTOMER CHURN PREDICTION**

The case data we are providing here 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.

**3.1 The 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. For example Hoekstra et al. defines a conceptual LTV model 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:

* 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.

**3.2 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. Hwang et al. defines the customer defection the hottest issue in highly competitive wireless telecom industry. 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.

Modelling customer churn in pure parametric perspective is not appropriate for LTV context because the retention function tends to be “spiky” and non-smooth, with spikes at the contract ending dates. And usually on the marketing perspective the sufficient information about the churn is the probability of possible churn. This enables the marketing department so that, given the limited resources, the high probability churners can be contacted first.

**Table1. Examples of churn prediction in literature**

|  |  |  |  |
| --- | --- | --- | --- |
| **Article** | **Market Sector** | **Case Data** | **Methods Used** |
| Au et al. | Wireless telecom | 100,000 subscribers | DMEL-method (data mining by evolutionary learning). |
| Buckinx et al. | Retail business | 158,884 customers | Logistic regression, ARD (automatic relevance determination), decision tree |
| Buckinx et al. | Daily grocery | 878 usable responses | MLR (multiple linear regression), ARD and decision tree |
| Ferreira et al. | Wireless telecom | 100,000 subscribers | Neural network, decision tree, hierarchial neuro-fuzzy systems, rule evolver |
| Gatland | Retail banking | 1,100 customers | Multiple regression |
| Hwang et al. | Wireless telecom | 16,384 subscribers | Logistic regression, neural network, decision tree |
| Mozer et al. | Wireless telecom | 46,744 subscribers | Logistic regression, neural network, decision tree |

Table 1 presents examples of the churn prediction studies found in literature. The methods used for churn analysis are presented in the table along with a case data size and market sector information. Buckinx et al. measures the loyalty and churn rate differently in retail setting. The loyal customers are those who shop frequently and at the same time exhibit a regular buying pattern. In this retail setting the customer churn is defined as customers who switch their purchases to another store. This is hard to detect because customers may still have transactions in the previous store. So Butnix et al. classify the customer a partial defective if he deviates from his established buying behavior . This is possible because in their setting they focus only on loyal clients.

Personal retail banking sector is a typical market sector where a customer is not regularly switching from one company to another. Customers usually give their banking business to one or two companies for long periods of time. This makes customer churn a priority for most companies in the banking sector. Garland has done research on customer profitability in personal retail banking. Although their main focus is on the customers’ value to the study bank, they also investigate the duration and age of customer relationship based on profitability. His study is based on customer survey by mail which helped him to determine the customer’s share of wallet, satisfaction and loyalty from the qualitative factors.

**4.TRADITIONAL METHODS OF SOLVING CUSTOMER CHURN**

Below are some traditional ways to reduce customer churn:

* **Analysing why churn occurs:** Yes, this may sound obvious, but let’s stress it once again: one needs to simply find out why customers decided to leave. The easiest way to do this is to talk to the customer.

In fact, [communicating with the customers](https://www.superoffice.com/blog/handling-customer-service-in-multiple-communication-channels-2/) does miracles in analysing churn. And one need to be actively using all channels for that: phone, e-mail, website, [live chat](https://www.superoffice.com/blog/live-chat-customer-experience/), and social media. The valuable feedback on how well you serve your customers is just a phone call, an e-mail or a survey away. As simple as that

* **Engaging with the customers**- Another way to prevent churn is to actively engage with customers about the company’s product.

Giving the customers reasons to keep coming back by showing them the day-to-day value of using the products, by making your products, services, offers, etc. a part of their daily workflow.

According to a recent [report from Marketo](https://www.marketo.com/analyst-and-other-reports/the-state-of-engagement/), the most efficient customer engagement channels for B2B companies to reach out to their existing customer base is through [email marketing](https://www.superoffice.com/blog/email-marketing-strategy/)**.**

* **Educating the customers**- This churn-prevention trick naturally flows from the point above. The company needs to provide enough good quality educational or support materials, which will help [increase retention](https://www.superoffice.com/blog/customer-retention-tips-with-crm-software/) and reduce churn. Offer free trainings, webinars, video tutorials, and product demos – whatever it takes to make customers feel comfortable and informed.

In other words, it is not only to give them the tools that work, but also offer the training on how to use these tools at a maximum profit. In this way the company will demonstrate the full potential of your products and services, and ensure that customers have [a successful onboarding and implementation](https://www.superoffice.com/blog/7-tips-to-a-successful-implementation-of-your-new-crm-system-2/).

* **Knowing who is at risk**-The best way to avoid churn is to prevent it from happening in the first place. There is always a group of customers that is more likely to leave than others – so it’s in your best interests to know who is balancing on that dangerous edge. This way you can reach out to them in time to make them stay.

Identifying at-risk customers is [one of the most popular](https://brightback.com/2020-brightback-state-of-industry) churn tactics for B2B companies. In fact, **35% of B2B of organizations have used this tactic to successfully reduce customer churn.**

* **Defining the most valuable customers-** A [history of the company’s interaction with the customers](https://www.superoffice.com/crm/service-features/) can show how deeply they are involved at each stage, whether they had any problems with the product, and whether these issues were dealt with.

So, what one can do is [segment your customers into groups](https://www.superoffice.com/blog/email-marketing-segmentation-mistake/) of profitability, readiness to leave, and their likelihood to positively respond to your offer to stay. In this way you can better predict customer churn.

* **Offering incentives**- Offering incentives and discount offers is widely regarded **as the most effective tactic in reducing churn.**

**5. PROPOSED METHOD**

**5.1 BINARY CLASSIFIER**

**Binary** or **binomial classification** is the task of [classifying](https://en.wikipedia.org/wiki/Statistical_classification) the elements of a given [set](https://en.wikipedia.org/wiki/Set_(mathematics)) into two groups (predicting which group each one belongs to) on the basis of a [classification rule](https://en.wikipedia.org/wiki/Classification_rule).

**Statistical Binary Classification:**

***Statistical Classification is a problem studied in machine learning. It is a type of***[***supervised learning***](https://en.wikipedia.org/wiki/Supervised_learning)***, a method of machine learning where the categories are predefined, and is used to categorize new probabilistic observations into said categories.*** When there are only two categories the problem is known as statistical binary classification.

Some of the methods commonly used for binary classification are:

* [Decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning)
* [Random forests](https://en.wikipedia.org/wiki/Random_forests)
* [Bayesian networks](https://en.wikipedia.org/wiki/Bayesian_network)
* [Support vector machines](https://en.wikipedia.org/wiki/Support_vector_machine)
* [Neural networks](https://en.wikipedia.org/wiki/Neural_network)
* [Logistic regression](https://en.wikipedia.org/wiki/Logistic_regression)
* [Probit model](https://en.wikipedia.org/wiki/Probit_model)

Each classifier is best in only a select domain based upon the number of observations, the dimensionality of the [feature vector](https://en.wikipedia.org/wiki/Feature_vector), the noise in the data and many other factors. For example, [random forests](https://en.wikipedia.org/wiki/Random_forests) perform better than [SVM](https://en.wikipedia.org/wiki/Support_vector_machine) classifiers for 3D point clouds.

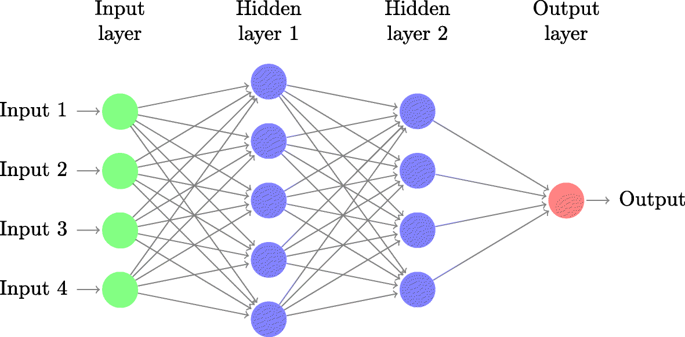
**In our model we have used neural networks as the method for binary classification.**

**5.2 ARTIFICIAL NEURAL NETWORK**

***Artificial neural networks****(****ANN****) or****connectionist systems are computing systems vaguely inspired by the***[***biological neural networks***](https://en.wikipedia.org/wiki/Biological_neural_network)***that constitute animal***[***brains***](https://en.wikipedia.org/wiki/Brain)***. Such systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules***

***An ANN is based on a collection of connected units or nodes called***[***artificial neurons***](https://en.wikipedia.org/wiki/Artificial_neuron)***.***

Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.



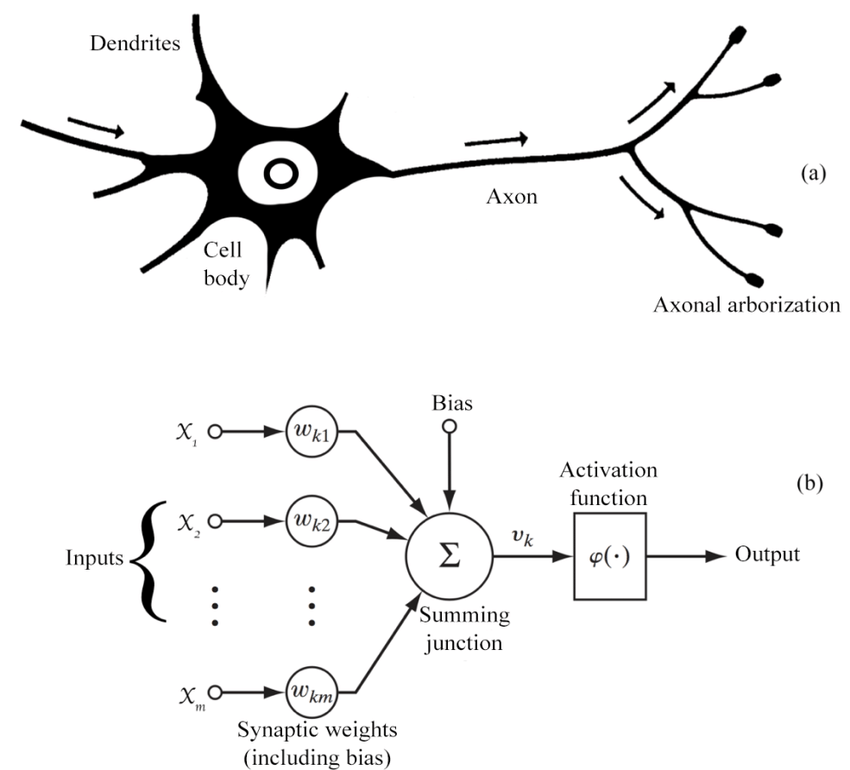


Fig. Diagram showing the structure of (a)animal neurons and (b)artificial neurons where x1,x2,…xn are the inputs wk1,wk2,wkm being the corresponding weights

5.2.1 Perceptrons and Multi-Layer Perceptrons (MLPs):

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), the **perceptron** is an algorithm for [supervised learning](https://en.wikipedia.org/wiki/Supervised_classification) of [binary classifiers](https://en.wikipedia.org/wiki/Binary_classification). A binary classifier is a function which can decide whether or not an input, represented by a vector of numbers, belongs to some specific class. It is a type of [linear classifier](https://en.wikipedia.org/wiki/Linear_classifier), i.e. a classification algorithm that makes its predictions based on a [linear predictor function](https://en.wikipedia.org/wiki/Linear_predictor_function) combining a set of weights with the [feature vector](https://en.wikipedia.org/wiki/Feature_vector).

The term "multilayer perceptron" refers to many perceptrons that are organized into layers. An alternative is "multilayer perceptron network". ***True perceptrons are formally a special case of artificial neurons that use a threshold activation function such as the***[***Heaviside step function***](https://en.wikipedia.org/wiki/Heaviside_step_function)***. MLP perceptrons can employ arbitrary activation functions.*** ***A true perceptron performs binary classification (either this or that), an MLP neuron is free to either perform classification or regression, depending upon its activation function.***

As we have used artificial neural networks here, so in the context of neural networks, a perceptron is an [artificial neuron](https://en.wikipedia.org/wiki/Artificial_neuron) using the [Heaviside step function](https://en.wikipedia.org/wiki/Heaviside_step_function) as the activation function. The perceptron algorithm is also termed the **single-layer perceptron**, to distinguish it from a [multilayer perceptron](https://en.wikipedia.org/wiki/Multilayer_perceptron), which is a misnomer for a more complicated neural network. As a linear classifier, the single-layer perceptron is the simplest [feedforward neural network](https://en.wikipedia.org/wiki/Feedforward_neural_network).

**5.2.2 Learning:**

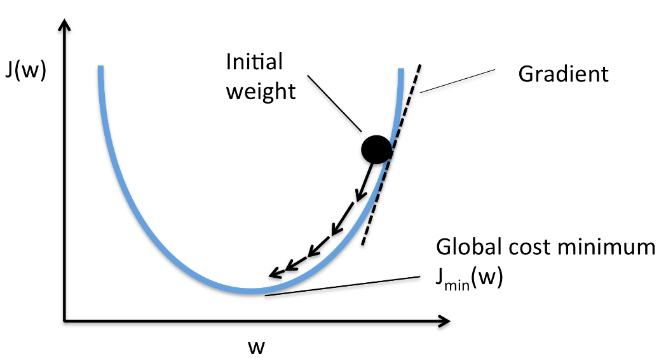
***Learning is the adaptation of the network to better handle a task by considering sample observations. Learning involves adjusting the weights (and optional thresholds) of the network to improve the accuracy of the result. This is done by minimizing the observed errors***.

**Learning is complete when examining additional observations does not usefully reduce the error rate.** The cost is frequently defined as a [statistic](https://en.wikipedia.org/wiki/Statistic) whose value can only be approximated.  Learning attempts to reduce the total of the differences across the observations. **The learning rate defines the size of the corrective steps that the model takes to adjust for errors in each observation.** The main objective for backpropagation is for distributing errors backward throughout the network to be used by neurons for adjusting individual weights.

In order for a neural network to learn, weights associated with neuron connections must be updated after forward passes of data through the network. These weights are adjusted to help reconcile the differences between the actual and predicted outcomes for subsequent forward passes. But how, exactly, do the weights get adjusted?

Before we get to the actual adjustments, what would be needed at each neuron in order to make a meaningful change to a given weight. Since we are talking about the difference between actual and predicted values, the **error** would be a useful measure here, and so each neuron will require that their respective error be sent backward through the network to them in order to facilitate the update process; hence, backpropagation of error. Updates to the neuron weights will be reflective of the magnitude of error propagated backward after a forward pass has been completed.

**Imagine that a cost function is used to determine our error (the difference between actual and predicted values), based on a given weight**. Consider the cost function below:



***Now, let's take as true the assertion that the lowest point on that cost function is the optimal value (minima), representing where the rate of change of the function is exactly zero.*** **Our objective is then to determine the value which produces this rate of change of zero.** How is this determined? Well, let's start somewhere on that function, with some value, and then use some method for determining where on the curve we are relative to the minima, which will then provide us with some clue as to what our next move should be, in order to make an attempt at reaching the bottom, where the rate of change is zero (which is optimal).

**Conceptually, the slope of the angle of our cost function at our current location can tell us if we are heading towards the right direction. As per basic algebra, a negative slope tells us we are headed downward (which is good), while a positive slope says that our previous step has overshot our goal (moved beyond the optimal and back up the other side of the function).**

But how to determine these slopes? As it turns out, **gradient** is actually a synonym for **derivative**, while derivative is the rate of change of a function. Well, that is exactly what we want. Descent indicates that we are spelunking our way to the bottom of a cost function using these changing gradients. And how do we get derivatives? By using the process of **differentiation**.

How far should we move in a direction towards the optimal i.e. the ***step size will have an effect on the how long it takes to reach the optimal value, how many steps it takes to get there, and how direct or indirect our journey is.***

So, what about Stochastic Gradient Descent (SGD)?

The process of gradient descent is very formulaic, in that it takes the entirety of a dataset’s forward pass and cost calculations into account in total, after which a wholesale propagation of errors backward through the network to neurons is made. This process would result in the same errors and subsequent propagated errors each and every time it is undertaken. Plain vanilla gradient descent is deterministic.

However, stochastic means randomly determined. **Instead of a rote processing of data, SGD uses a random sampling of the data to perform the same steps which are performed via a full set of data in vanilla gradient descent. This can speed up learning, as well as lead to different (possibly better) results over a number of iterations**.

5.3 CONFUSION MATRIX AND AUC-ROC CURVE

**Confusion Matrix comes into the picture as a metrics when measuring the effectiveness or performance of our model. We will analyse the prediction results of our classification model by using the confusion matrix.**

It is a table with 4 different combinations of predicted and actual values.



It is extremely useful for measuring **Recall, Precision, Specificity, Accuracy and most importantly AUC-ROC Curve.**

Let us understand the analogy of TP, FP, TN, FN:

* **TP (or True Positive) is that the model predicted positive and it actually is.**
* **FP ( or False Positive) is that the model predicted false and it actually is.**
* **TN (or True Negative) is that the model predicted true but it is actually false.**
* **FN ( or False Negative) is that the model predicted false but it is actually true.**

**RECALL**- Out of all the positive classes (“churn” class), how much the model predicted correctly. It should be as high as possible.

**Recall = TP / TP + FN**

**PRECISION**- Out of all the positive classes that the model predicted correctly, how many are actually positive.

**Precision = TP / TP + FP**

**ACCURACY**- Out of all the classes (including both “churn” and “no-churn”), how much the model predicted correctly. It should be as high as possible.

**Accuracy = TP + TN / Total of Predicted and Actual Values**

**F-SCORE** – It is difficult to compare two models with low precision and high recall or vice versa. So, to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.

**F-Score = 2 \* Recall \* Precision / Recall + Precision**

AUC-ROC (Area Under Curve-Receiver Operating Characteristics) Curve:

AUC - ROC curve is a performance measurement metrics for classification problem at various threshold settings. ***ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between churn and no churn. It is also written as Area Under Receiver Operator Characteristics (AUROC).***

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

**Terms used in AUC-ROC Curve:**

**TPR (True Positive Rate) / Recall / Sensitivity calculated as the number of correct positive predictions divided by the total number of positives =** TP / TP + FN

**Specificity (True Negative Rate) calculated as the number of correct negative predictions divided by the total number of negatives =** TN / TN + FP

**FPR (False Positive Rate)** = 1 – Specificity = FP / TN + FP

**How to speculate the performance of the model?**

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact, it means, it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

**6.RESULTS AND DISCUSSIONS**

6.1 Case Data:

Our dataset consists of 10,000 rows and 14 columns. The index starts from 0 to 999 and the columns are as follows: CustomerID, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited.



Fig. The entire dataset consisting of 10,000 rows and 14 columns

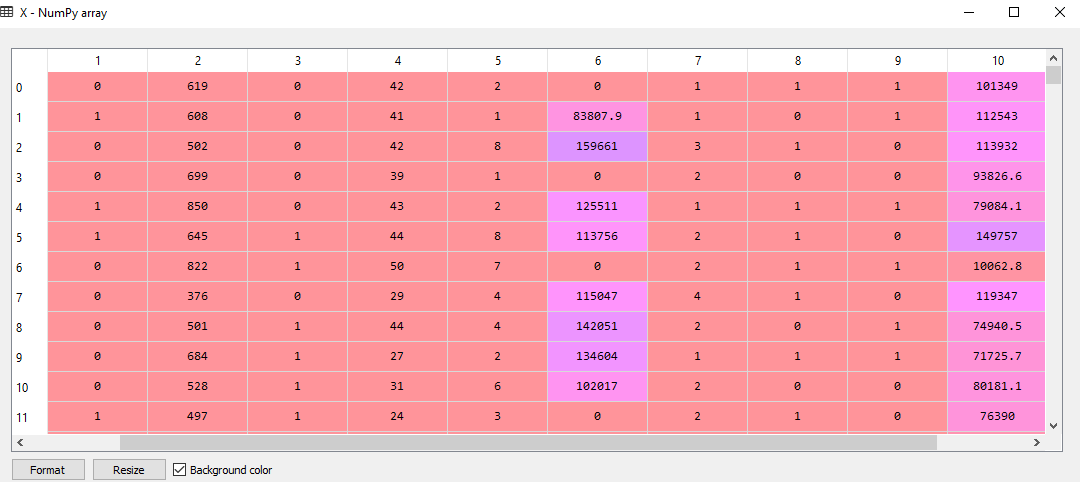
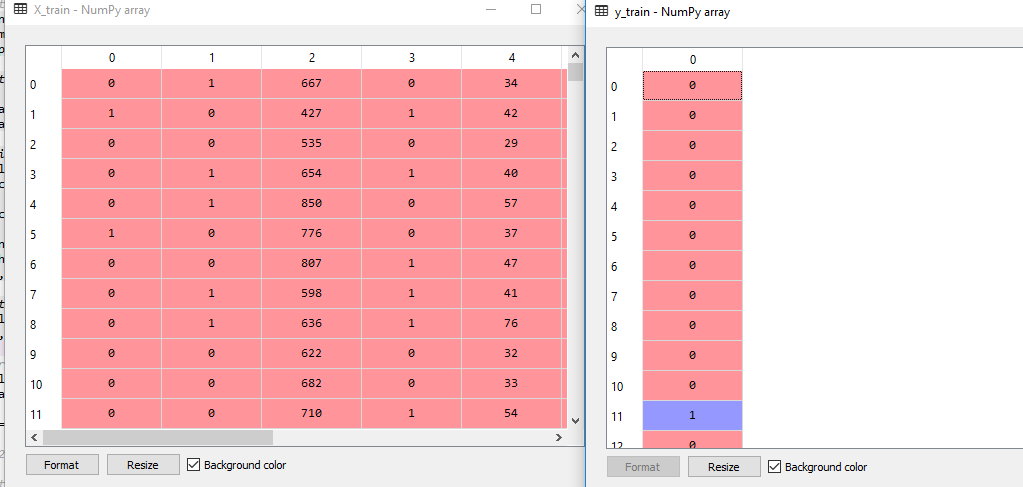


Fig. The feature matrix ‘X’ consisting of dependent variables



Fig. The Dependent Matrix ‘y’



**Fig. The training sets ‘X\_train’ and ‘y\_train’**

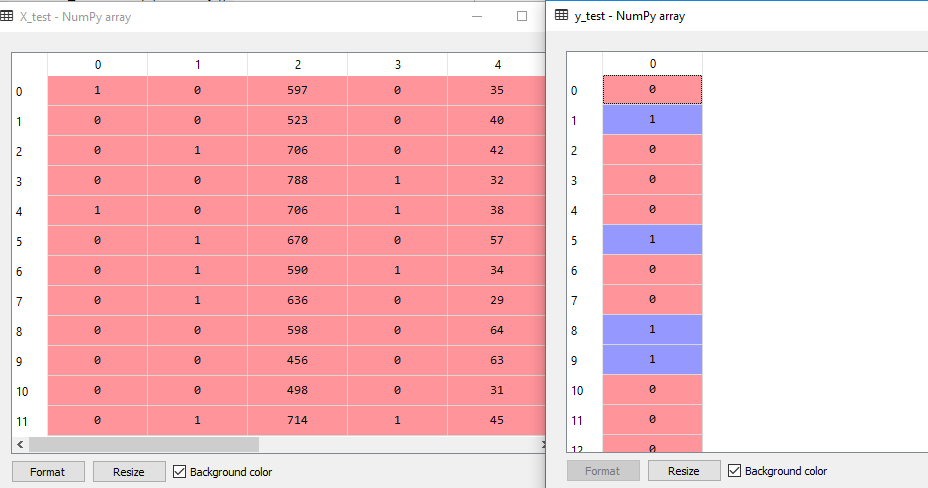


Fig. The testing sets ‘X\_test’ and ‘y\_test’

**6.2 Case Focus:**

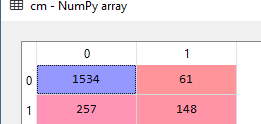
The main focus of this case study is to see how well the model predicts the current scenario of customer churn as presented in our case data which would give us a hint that whether a binary classifier along with neural network can be used for predicting customer churn in the future.

We would calculate the recall, precision and most importantly the accuracy of our model from the confusion matrix and plot the AUROC curve to evaluate our model’s performance more accurately.

**6.3 Predictive Performance and Model Evaluation:**

**6.3.1** **CONFUSION MATRIX:**

We have the following confusion matrix:



The interpretations from the above confusion matrix are as follows—

* **The model predicted that 1534 people are present in bank which is actually true.**
* **The model predicted that 148 people have left the bank and is also true.**
* **The model predicted that 61 people are present in the bank which is false in accordance with the original results.**
* **The model predicted that 257 people have left the bank which is also false**

We have

**Recall** = TP / TP + FN = 1534 / 1534 + 257 = 0.85

**Precision** = TP / TP + FP = 1534 / 1534 + 61 = 0.96

**Accuracy**

= TP + TN / Actual values + Predicted values = 1534+148/1595+405= 11682/2000= 0.84 ≈ 84 %

**F-Score** = 2 \* 0.85 \* 0.96 / 0.85 + 0.96 = 1.63 / 1.81 = 0.90

A classifier with a precision of 1.0 and a recall of 0.0 has a simple average of 0.5 but an F1 score of 0. The F1 score gives equal weight to both measures and is a specific example of the general Fβ metric where β can be adjusted to give more weight to either recall or precision. (There are other metrics for combining precision and recall, such as the [Geometric Mean of precision and recall](https://en.wikipedia.org/wiki/Fowlkes%E2%80%93Mallows_index), but the F1 score is the most commonly used.) If we want to create a balanced classification model with the optimal balance of recall and precision, then we try to maximize the F1 score.

**6.3.2. AUC-ROC Curve:**

**The AUC-ROC metric is used to evaluate a classifier output quality**. **ROC curve (Receiver Operating Characteristic)** is a commonly used way to visualize the performance of a binary classifier and **AUC (Area Under the ROC Curve) is used to summarize its performance in a single number.**

Most machine learning algorithms have the ability to produce probability scores that tells us the strength in which it thinks a given observation is positive. Turning these probability scores into yes or no predictions requires setting a threshold; a default might be to use a threshold of 0.5, meaning that a probability in [0.0, 0.49] is a negative outcome (0) and a probability in [0.5, 1.0] is a positive outcome (1).i.e. cases with scores above the threshold are classified as positive, and vice versa. Different threshold values can lead to different result:

* A higher threshold is more conservative about labelling a case as positive; this makes it less likely to produce false positive (an observation that has a negative label but gets classified as positive by the model) results but more likely to miss cases that are in fact positive (lower true positive rate)
* A lower threshold produces positive labels more liberally, so it creates more false positives but also generate more true positives

ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the “ideal” point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better.

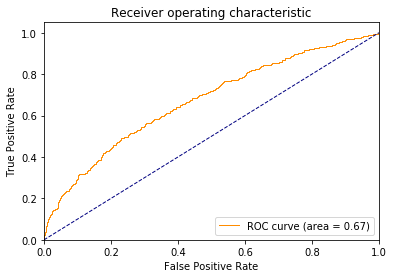


Fig. AUC-ROC curve of the model

**Here, as the graph above shows we are getting AUROC value of 0.67.**

The “steepness” of ROC curves is also important, since it is ideal to maximize the true positive rate while minimizing the false positive rate.

ROC curves are typically used in binary classification to study the output of a classifier.

Now,

**True Positive Rate (which is also known as the Recall or Sensitivity)**

= No. of positive data points with positive predictions/ No. of all positive data points

= TP / TP + FN = 0.85

**False Positive Rate**

= No. of negative data points with positive predictions/ No. of all negative data points

= FP / FP + TN = 0.19

**7.FUTURE SCOPE AND CONCLUSION**

In this report, a customer churn analysis was presented in a retail personal banking sector. The analysis focused on churn prediction based on binary classification with artificial neural network. The model predicted the actual churners relatively well. The differences between the model’s input data (the significance level in case of each of the variables) indicates the dynamic nature of the churning customer profile. This makes it hard to formulate one standard model that could be used as the predictive model in the future. The findings of this study indicate that, in case of neural network model, the user should update the hyperparameters of the network to be able to produce predictions with high accuracy.

It is interesting for a company’s perspective whether the churning customers are worth retaining or not and also in marketing perspective what can be done to retain them.

The customer churn analysis in this study might not be interesting if the customers are valued based on the customer lifetime value. The churn definition in this study was based on the current account. But if the churn definition was based on for example loyalty program account or active use of the internet service. Then the customers at focus could possibly have greater lifetime value and thus it would be more important to retain these customers.

**8. REFERENCES**

* <https://www.vttresearch.com/sites/default/files/julkaisut/muut/2006/customer_churn_case_study.pdf>
* <https://www.kdnuggets.com/2017/10/neural-network-foundations-explained-gradient-descent.html>
* <https://medium.com/@pushkarmandot/build-your-first-deep-learning-neural-network-model-using-keras-in-python-a90b5864116d>
* <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
* <https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/>
* <https://ethen8181.github.io/machine-learning/model_selection/auc/auc.html>
* <https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html>
* <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>