Customer Churn Prediction For Banking Sector

Submitted by :-
HARMAN BHUTANI

Agenda

- Introduction : Customer Churn & Use in banking
- Business Challenges & Project Objective
- System Overview
  - Data Setup
  - Pre-Analytics
  - Model Development
  - Evolution
  - Execution
- Results
- Conclusion

Customer Churn

Customer churn is when an existing customer, user, player, subscriber or any kind of return client stops doing business or ends the relationship with a company

Customer churn has become a major problem within a customer centred banking industry and banks have always tried to track customer interaction with the company, in order to detect early warning signs in customer's behaviour such as reduced transactions, account status dormancy and take steps to prevent churn

Churn rate usually lies in the range from 10% upto 30%

Churn Prediction Importance In Banking

CONTRACTUAL

- Customers make purchases at discrete intervals, on a contract or autopay
- Cancellation event is observed and recorded
- Example: Netflix, cell phone service provider

NON-CONTRACTUAL

- Customers are free to buy or not at any time
- Churn event is not explicitly observed
- Example: Online fashion retailer

VOLUNTARY

- Customers make the choice to leave the service

INVOLUNTARY

- Customers are forced to discontinue service and/or payments
- Example: credit card expiration

Project Objective

Project Objective	The object to show customer churn process for a bank and how the business analytics will work in predicting those churn customers and benefits the bank
Business Objective	For the bank, objective are to gain insights from its past data, and to identify customers any stage of their lifecycle who are currently active but are likely to become inactive
Model Objective	The model would rank each customer between 0 and 1 on the basis of their probability to churn based on the 6 months historical behaviour
Target Base	Total data of 10,000 customers with 14 columns

Approach For Churn Prediction

Data Setup	Pre-Analytics	Model Development	Evaluation	Execution
Collect the right data and set up ADS - analytical datastore for modeling	Identify the main static, demographic and behaviour levers and their influences on dormancy to understood correlation based on available data	Develop a propensity model to rank the subscribers based on their likelihood to churn/ become inactive	Comparative evaluation of various modeling techniques and the best performing model will be selected	Design retention campaigns Set up test and control groups Execute and measure results from campaign

Data Setup

Pre-Analytics

Model Development

Evaluation

Execution

Data Information

Number	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

- Above data set of the bank showed the variables over which we need to perform the data analytics application to find out the churn customers

- 70% of the time, the data scientist is busy working on setting up the data, cleansing the data and profiling the data for feature engineering

Features

Features	Details	Features	Details
RowNumber	Index no. of row	Tenure	How long customer with bank
CustomerId	Customer ID	Balance	Current balance in account
Surname	Last name of customer	NumOfProducts	Prod. taken by cust.
CreditScore	Credit score given by bank	HasCrCard	Owning credit card or not
Geography	Country of customer	IsActiveMember	Active or not
Gender	Gender of customer	EstimatedSalary	Annual salary of customer
Age	Age of customer	Exited	Cust. still with bank or not

Categorical & Numerical Features

<u>Categorical Features</u>

Variable that are fixed or limited number of possible values assigning each individual or other unit of observations

Geography	France,germany, Spain
Gender	Male, Female
NumOfProducts	1, 2, 3, 4
HasCrCard	0 = No, 1 = Yes
IsActiveMember	0 = No, 1 = Yes

<u>Numerical Features</u>

These variables are numerical and have meaning as measurement

CreditScore	Numerical value given by bank
Age	Customer age
Tenure	Using services
Balance	Current balance
EstimatedSalary	Annual salary

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
RowNumber	10000.0	5.000500e+03	2886.895680	1.00	2500.75	5.000500e+03	7.500250e+03	10000.00
CustomerId	10000.0	1.569094e+07	71936.186123	15565701.00	15628528.25	1.569074e+07	1.575323e+07	15815690.00
CreditScore	10000.0	6.505288e+02	96.653299	350.00	584.00	6.520000e+02	7.180000e+02	850.00
Age	10000.0	3.892180e+01	10.487806	18.00	32.00	3.700000e+01	4.400000e+01	92.00
Tenure	10000.0	5.012800e+00	2.892174	0.00	3.00	5.000000e+00	7.000000e+00	10.00
Balance	10000.0	7.648589e+04	62397.405202	0.00	0.00	9.719854e+04	1.276442e+05	250898.09
NumOfProducts	10000.0	1.530200e+00	0.581654	1.00	1.00	1.000000e+00	2.000000e+00	4.00
HasCrCard	10000.0	7.055000e-01	0.455840	0.00	0.00	1.000000e+00	1.000000e+00	1.00
IsActiveMember	10000.0	5.151000e-01	0.499797	0.00	0.00	1.000000e+00	1.000000e+00	1.00
EstimatedSalary	10000.0	1.000902e+05	57510.492818	11.58	51002.11	1.001939e+05	1.493882e+05	199992.48
Exited	10000.0	2.037000e-01	0.402769	0.00	0.00	0.000000e+00	0.000000e+00	1.00

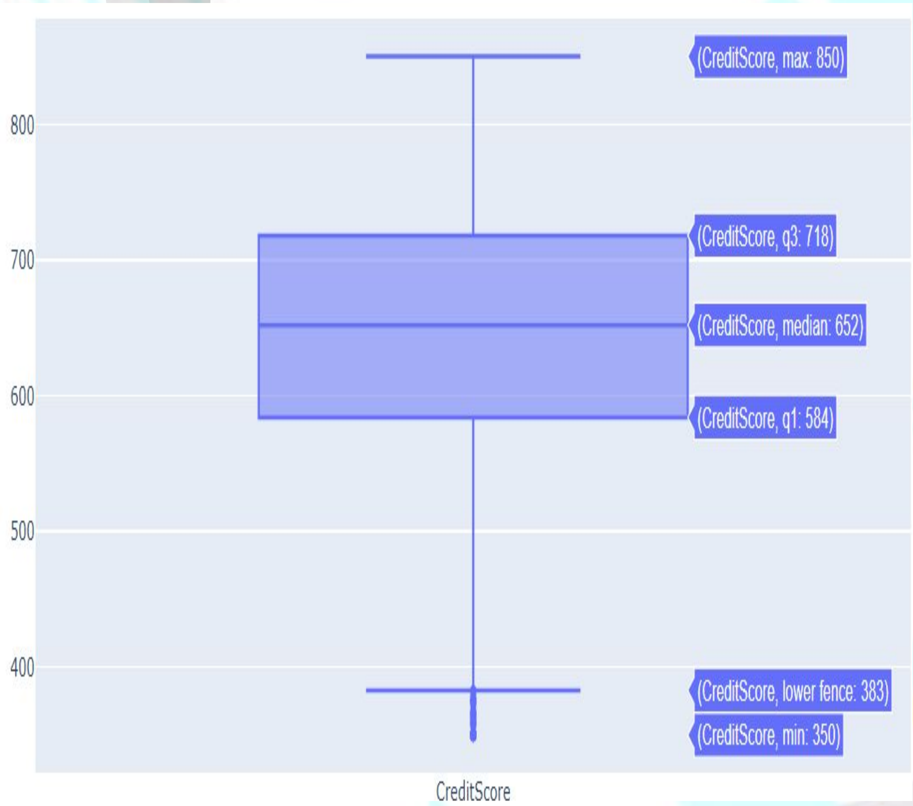
Outliers Detection

Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty.

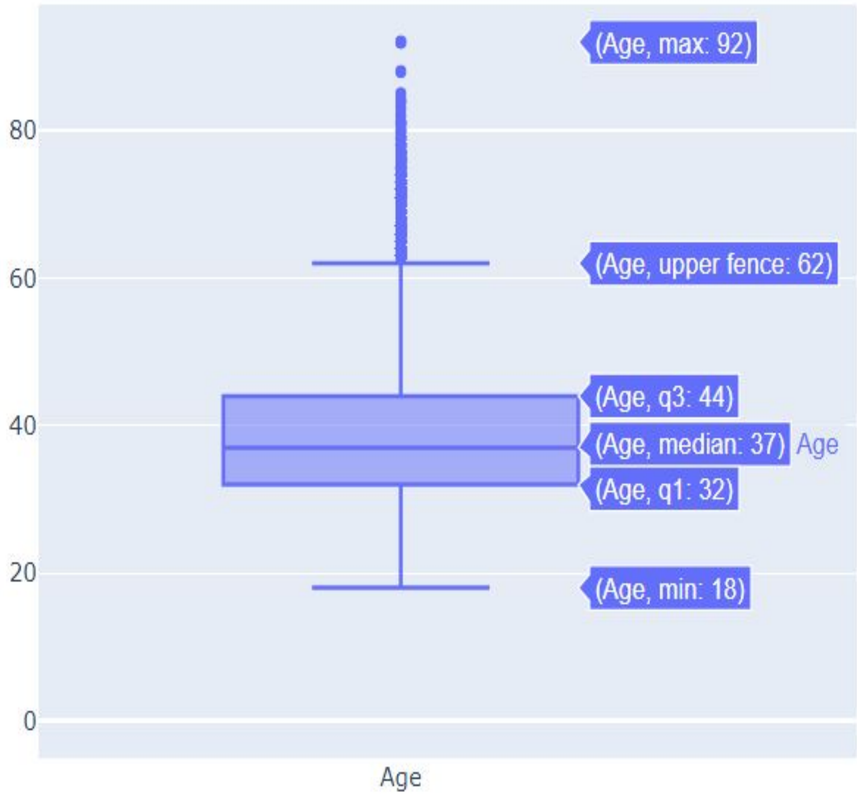
Box Plot

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“**Minimum**”, **First Quartile (Q1)**, **Median (Q2)**, **IQR, Third Quartile (Q3)**, and “**Maximum**”)

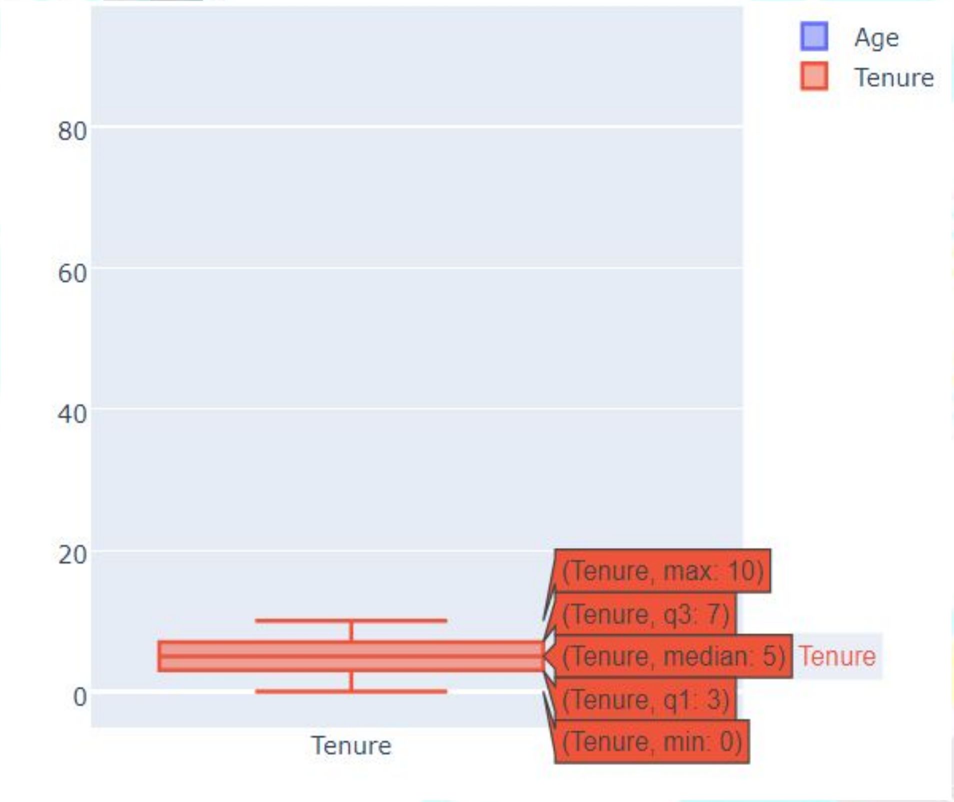
CreditScore



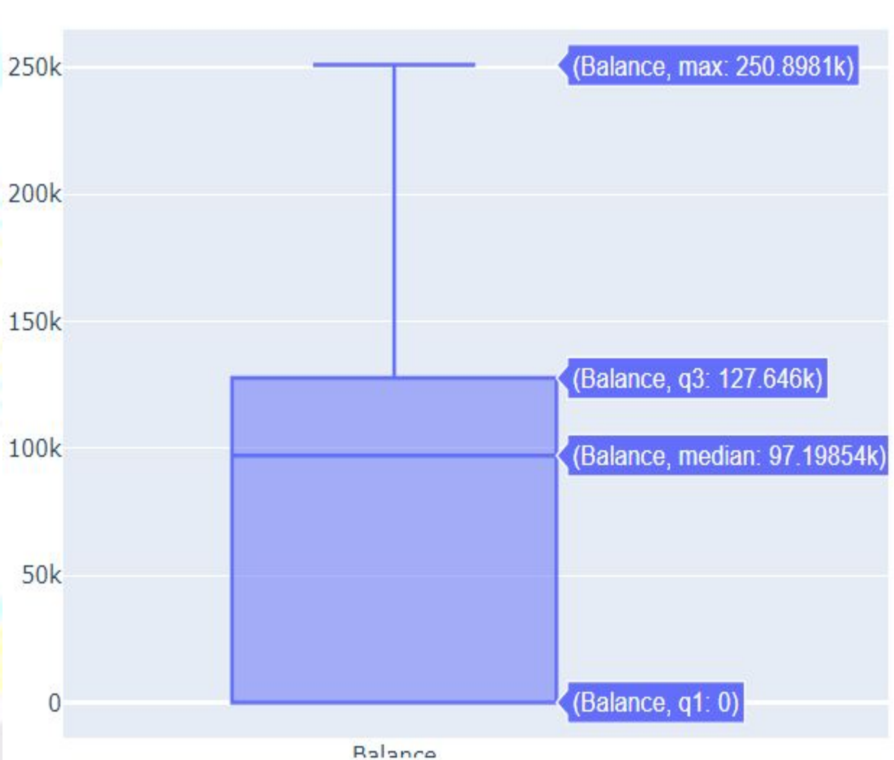
Age



Tenure



Balance



Features Scaling

- **Feature scaling** is a method used to normalize the range of independent variables or features of data

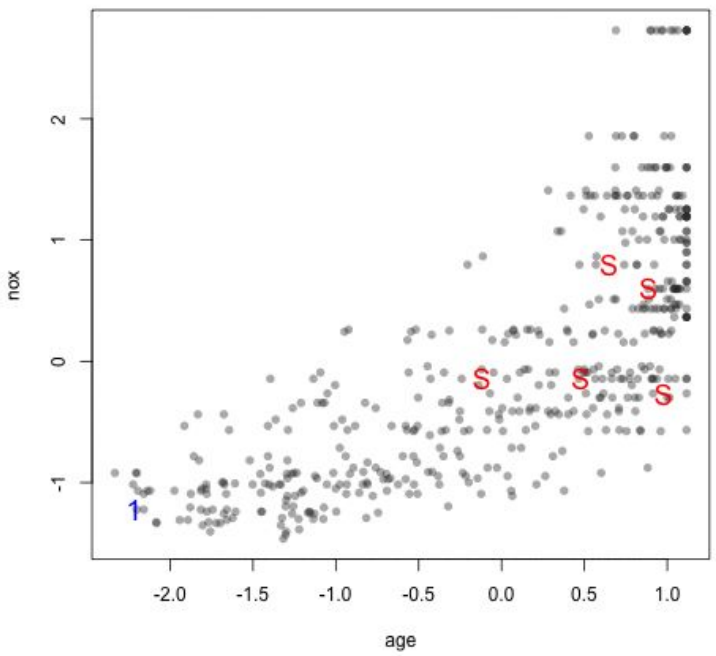
- We have normalize the dataset using standardization mechanism

- Feature **standardization** makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance.

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where x is the original feature vector, $\bar{x} = \text{average}(x)$ is the mean of that feature vector, and σ is its standard deviation.

Data Splitting (Train, Validation, Test)



- Dataset is divided into three groups Training dataset, Validation dataset and Test dataset
- Total Samples= 10,000
- Training Samples = 7,000
- Validation Samples = 1,000
- Testing Samples = 2,000

Data Setup

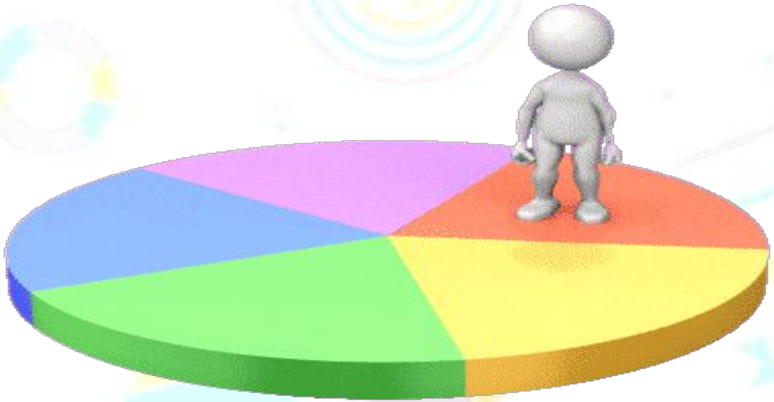
Pre-Analytics

Model Development

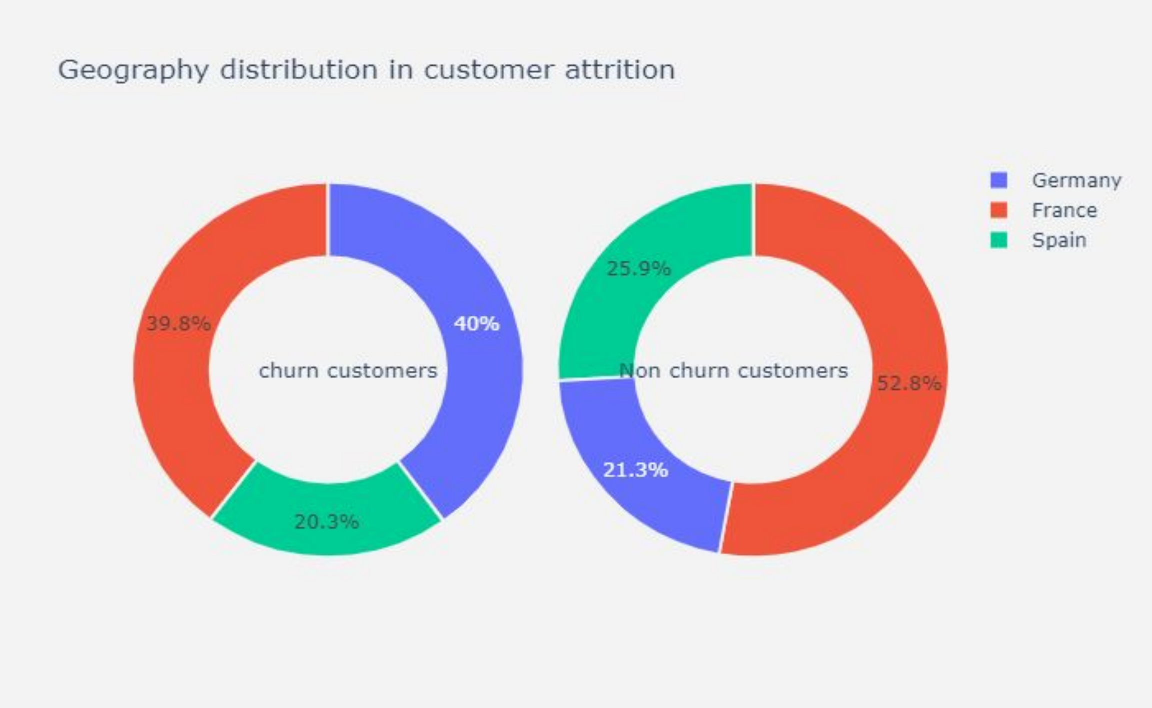
Evaluation

Execution

Categorical Data Visualization

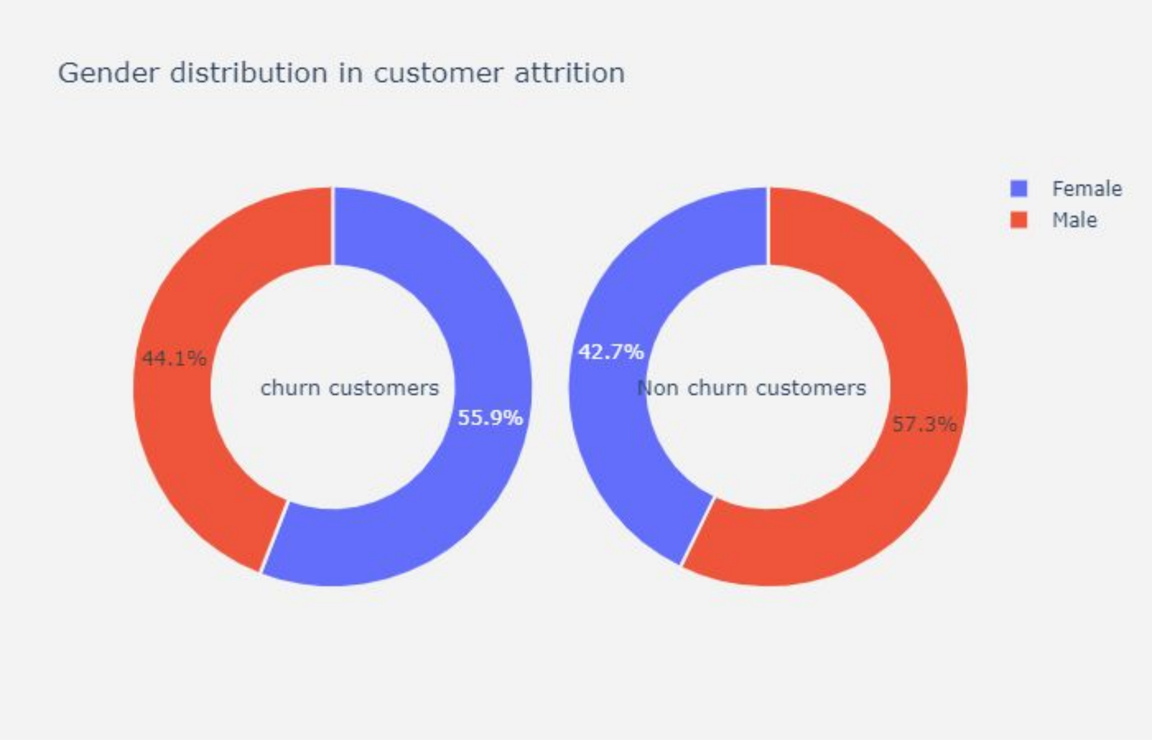


Geography



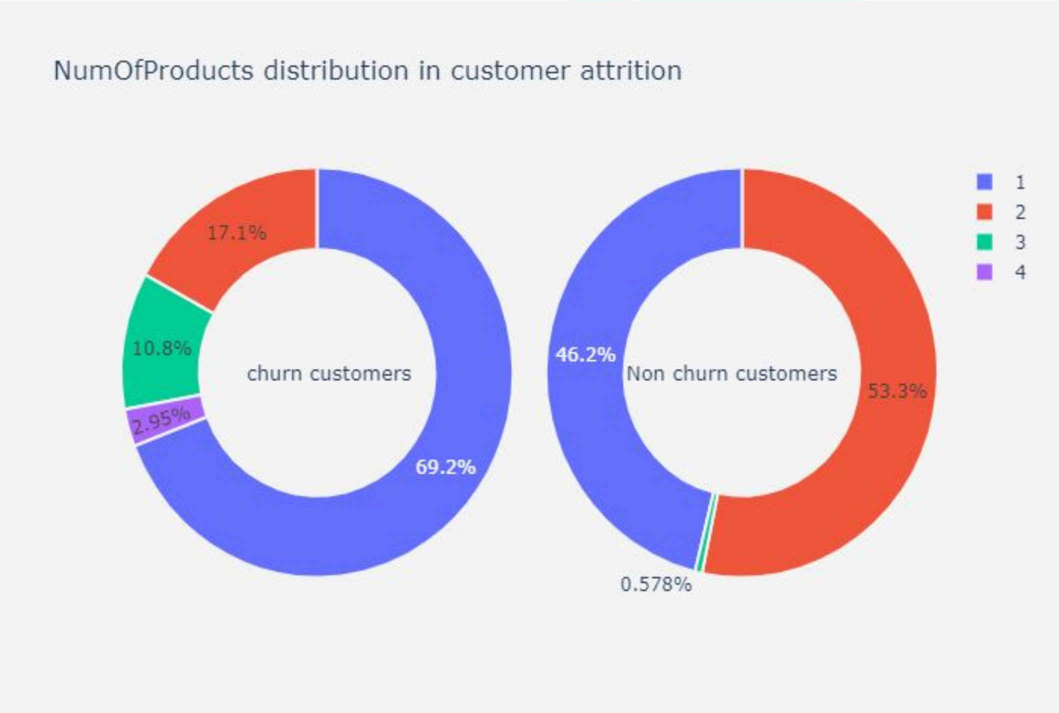
- Division of data on the basis of categorical variable Geography.
- These 2 pie chart shows the number of customer in each country divided as churn customers and non-churn customers
- Used to compare the churn and non-churn ratio as per the country

Gender



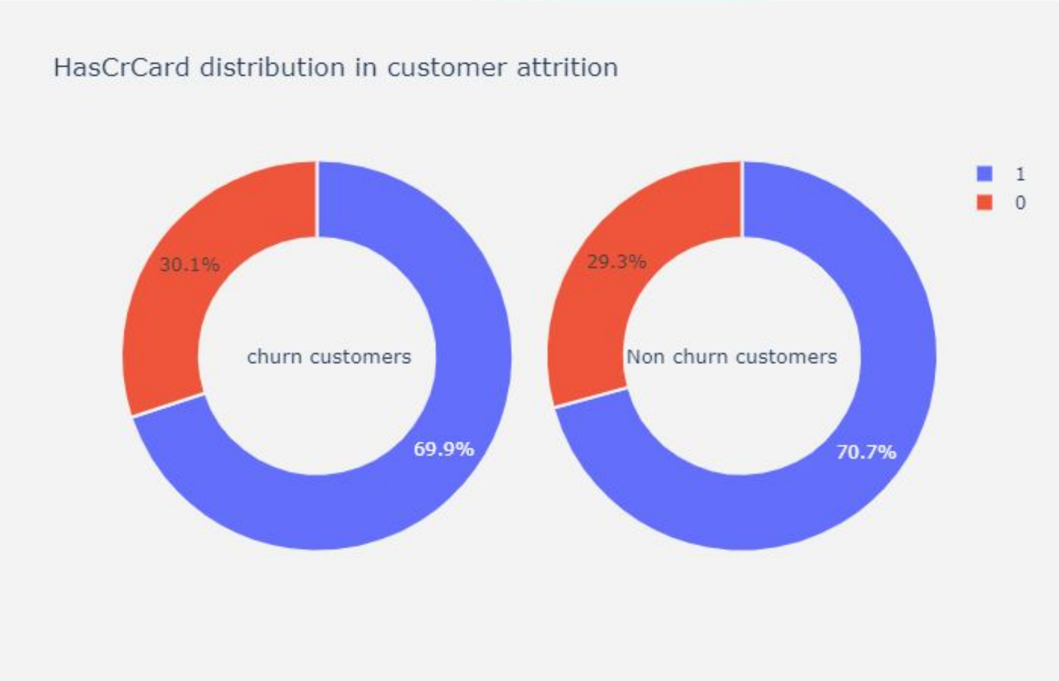
- Division of data on the basis of categorical variable gender
- These 2 pie chart shows the number of customer divided as churn customers and non-churn customers of the basis of gender
- This will give the insight view of customers on the basis of gender

Number of Products



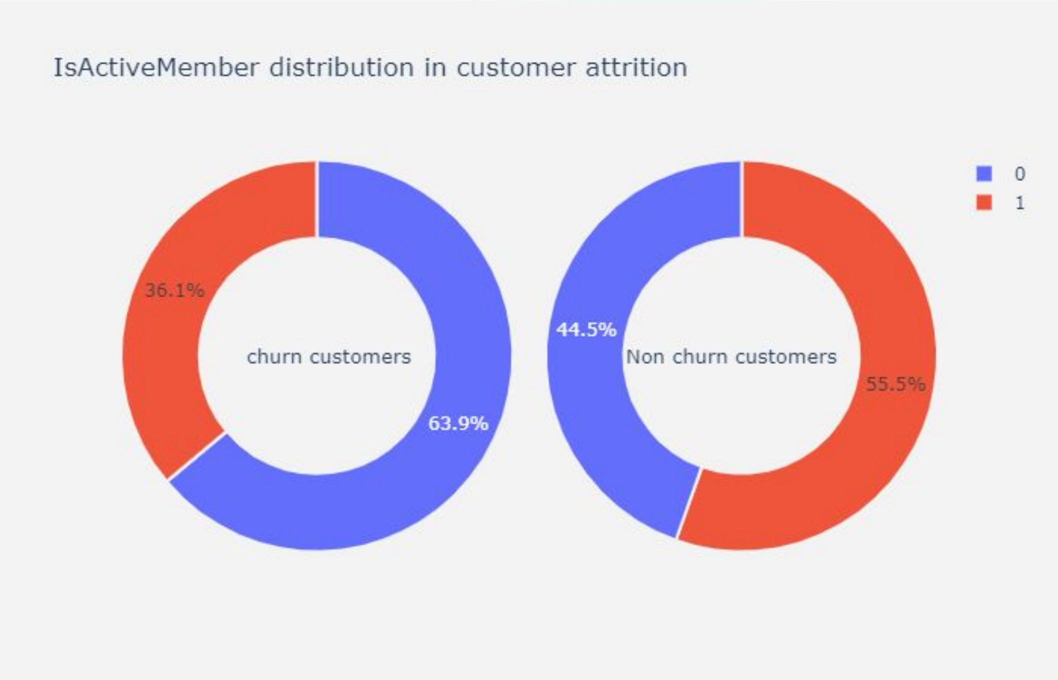
- Division of data on the basis of categorical variable number of products
- These 2 pie chart shows the number of customer divided as churn customers and non-churn customers of the basis of number of products/services that are used by those customers
- This will give us the ratio between how much services used and they churn

Has Credit Card



- Division of data on the basis of categorical variable HasCrCard which shows customer has a credit card or not
- These 2 pie chart shows the number of customer divided as churn customers and non-churn customers of the basis of credit card whether they have it or not
- This will give us the ratio between customer churn which has credit card or not having it

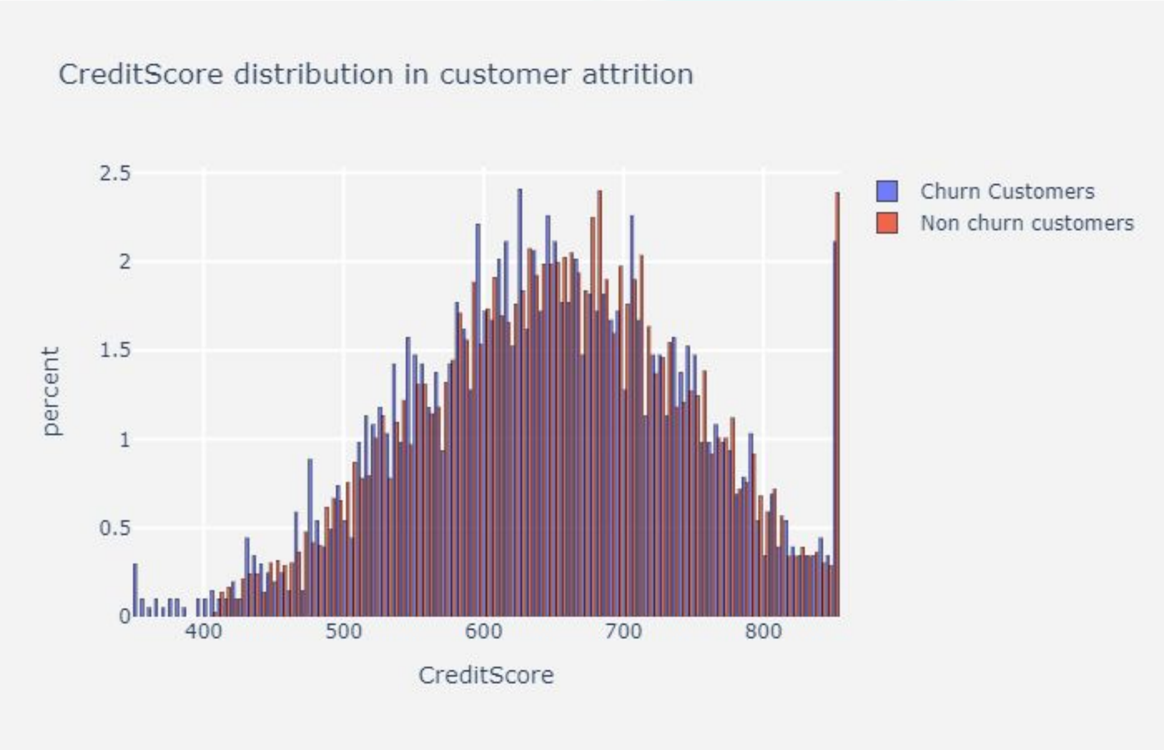
Is Active Member



- Division of data on the basis of categorical variable IsActiveMember
- These 2 pie chart shows the number of customer divided as churn customers and non-churn customers of the basis of whether the customer is active member or not
- This will give us the ratio between customer churn which are active or not

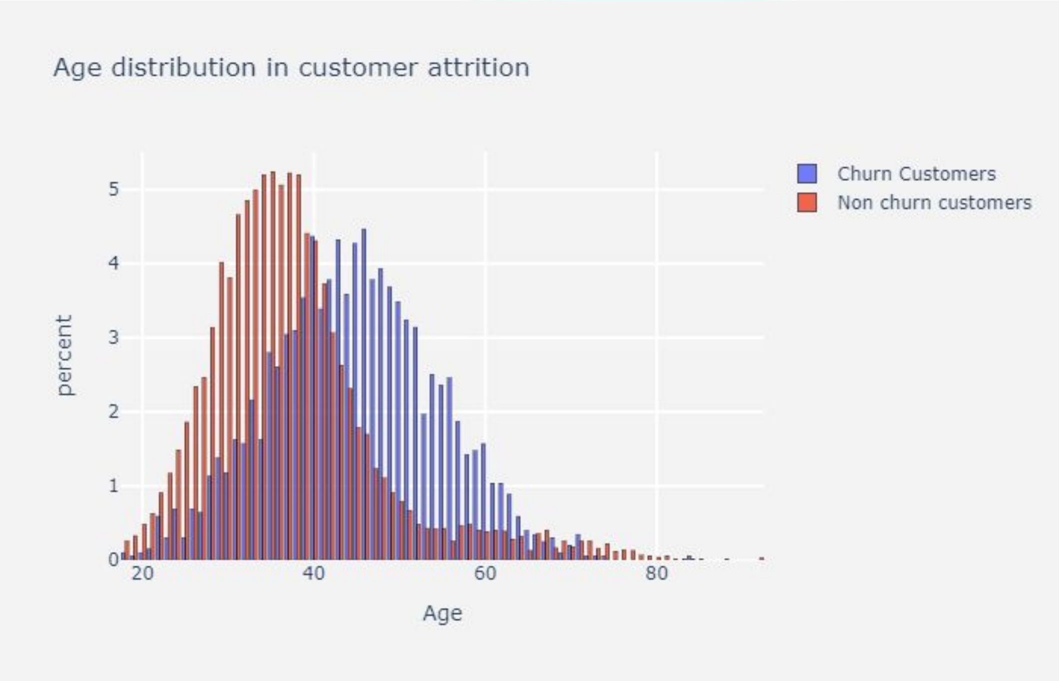
Numerical data Visualization

Credit Score



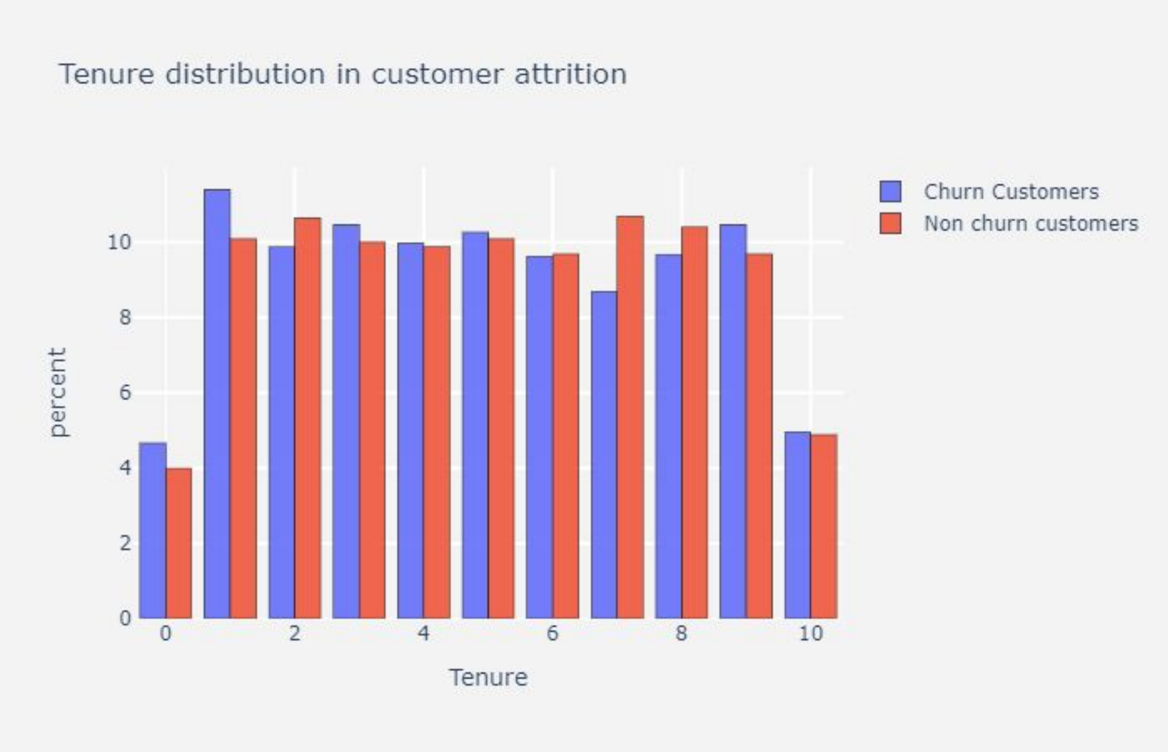
- Division of data on the basis of Numerical variable Credit Score
- Histogram will show the customers that are churn or non churn on the basis of their credit scores
- This will give us the ratio between customer churn which are having high credit score or low credit score
- Mean(Churn) : 625-629
- Mean(Non churn) : 680-684

Age



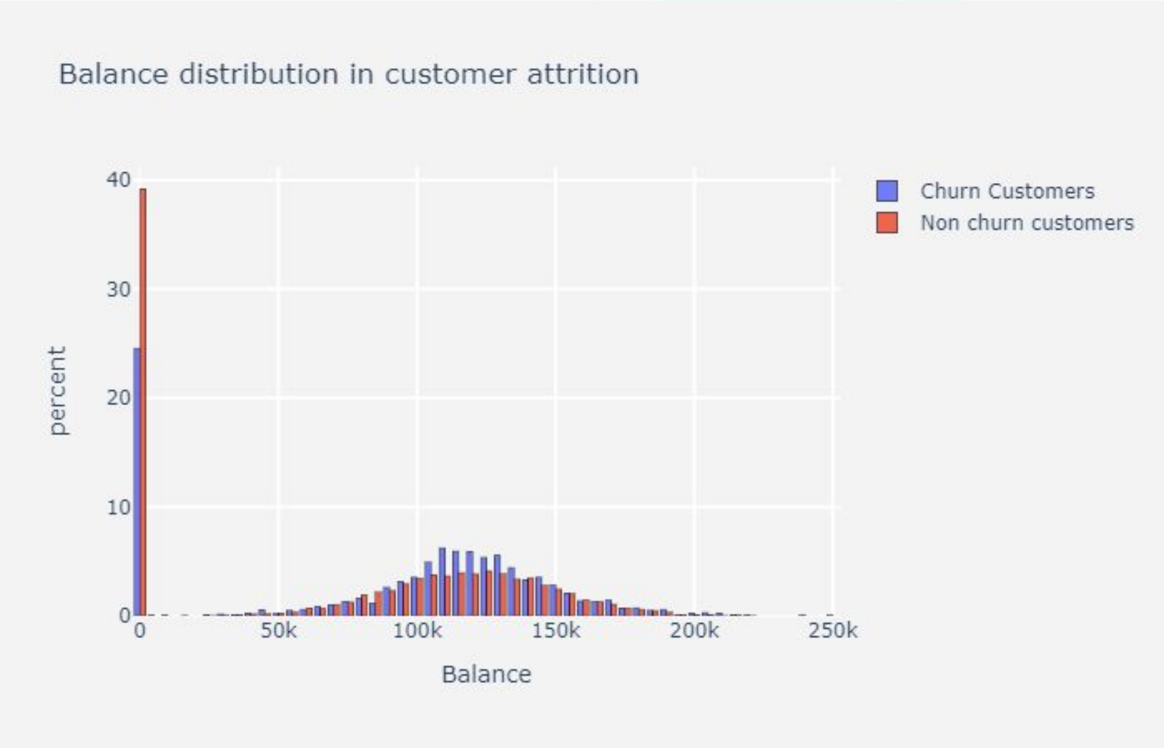
- Division of data on the basis of Numerical variable Age
- Histogram will show the customers that are churn or non churn on the basis of their Age
- This will give us the ratio between customer churn Age and non churn customers age
- Mean(Churn) : 46
- Mean(Non churn) : 35

Tenure



- Division of data on the basis of Numerical variable Tenure
- Histogram will show the customers that are churn or non churn on the basis of their Tenure
- This will give us the ratio between customer churn that are using services from how much time
- This will give an idea about after how much time the customers are usually churn

Balance



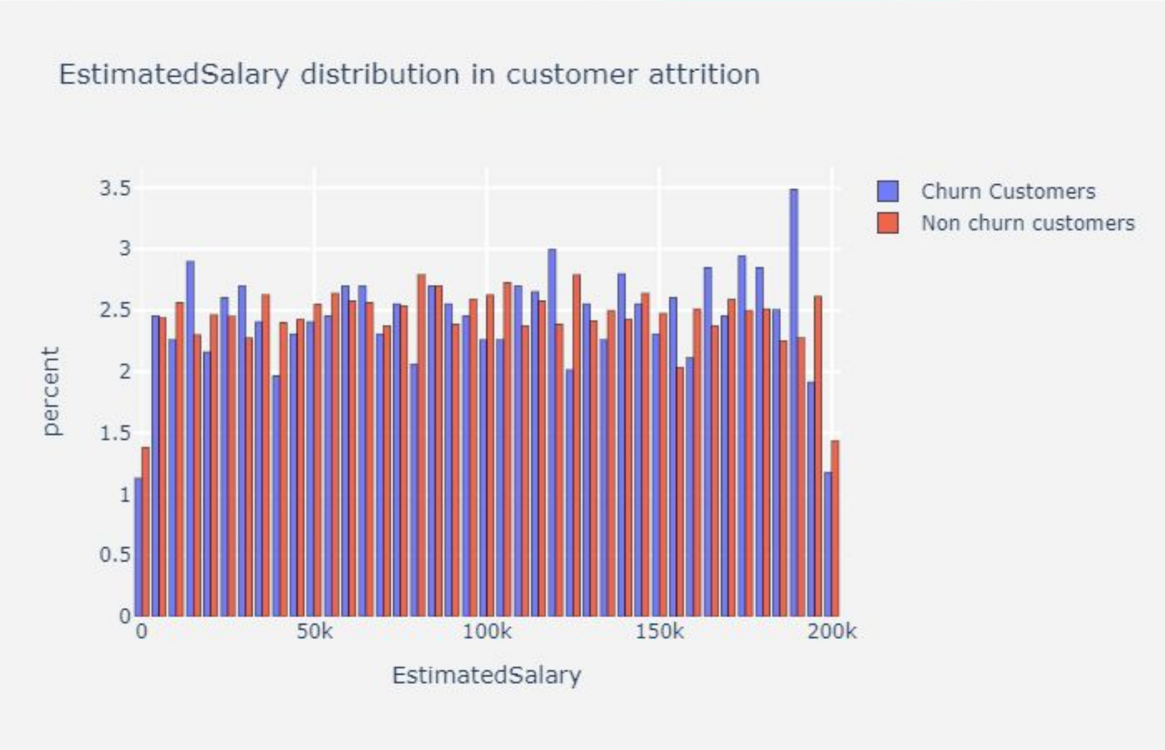
- Division of data on the basis of Numerical variable Balance

- Histogram will show the customers that are churn or non churn on the basis of their Balance available in account

- This will give us the ratio between customer churn having balance in account or not

- Mean(Churn) : 107.5K - 112.5K
- Mean(Non churn) : 112.5K - 117.5K

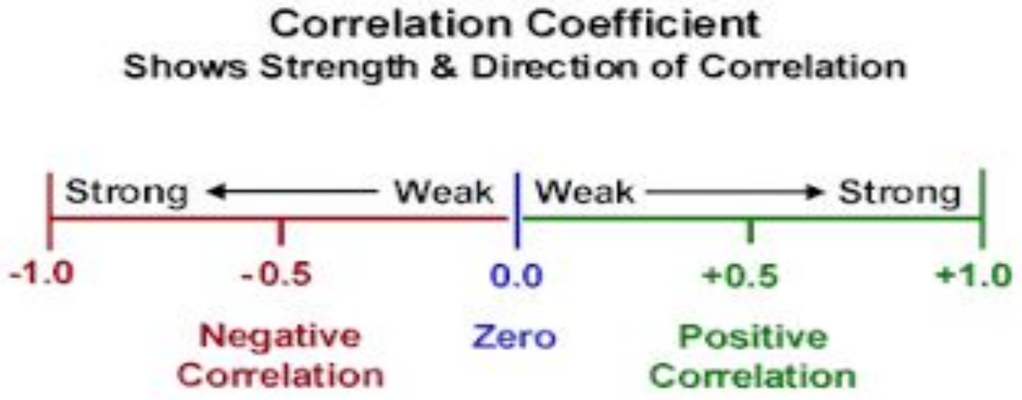
Estimated Salary



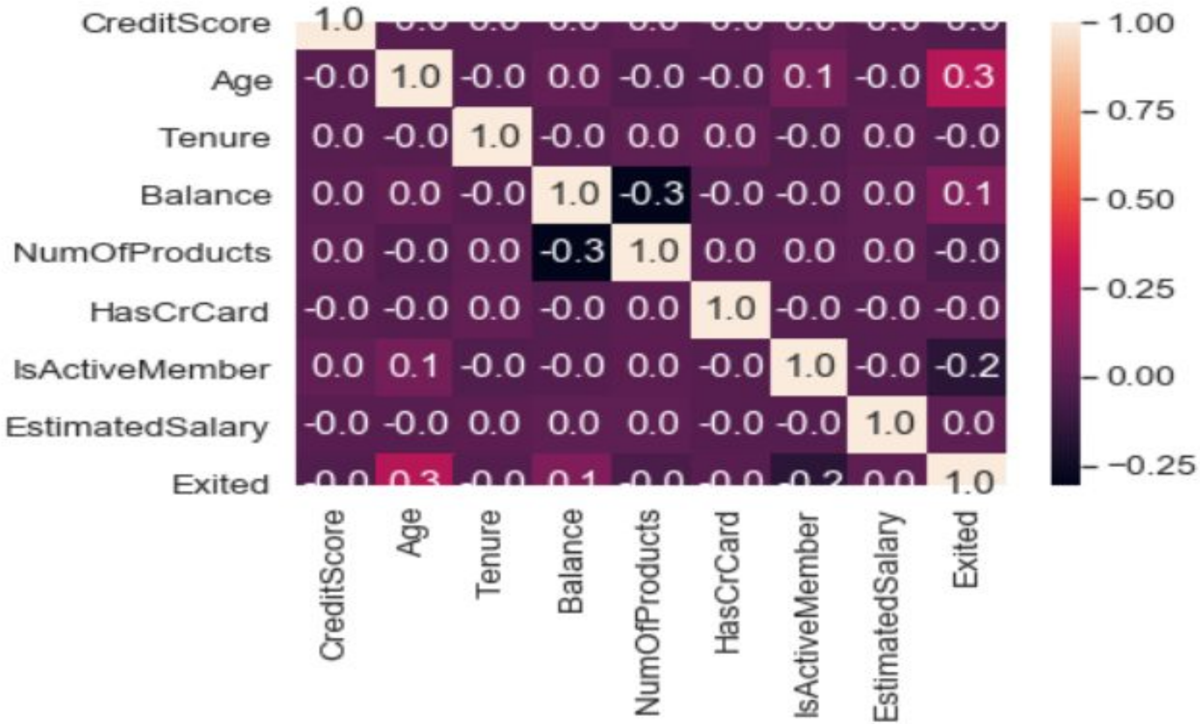
- Division of data on the basis of Numerical variable Salary
- Histogram will show the customers that are churn or non churn on the basis of their Salary which will be credit in account
- This will give us the ratio between customer churn having salary account or not
- This will give us an idea about whether churn depends on the salary or not

Correlation

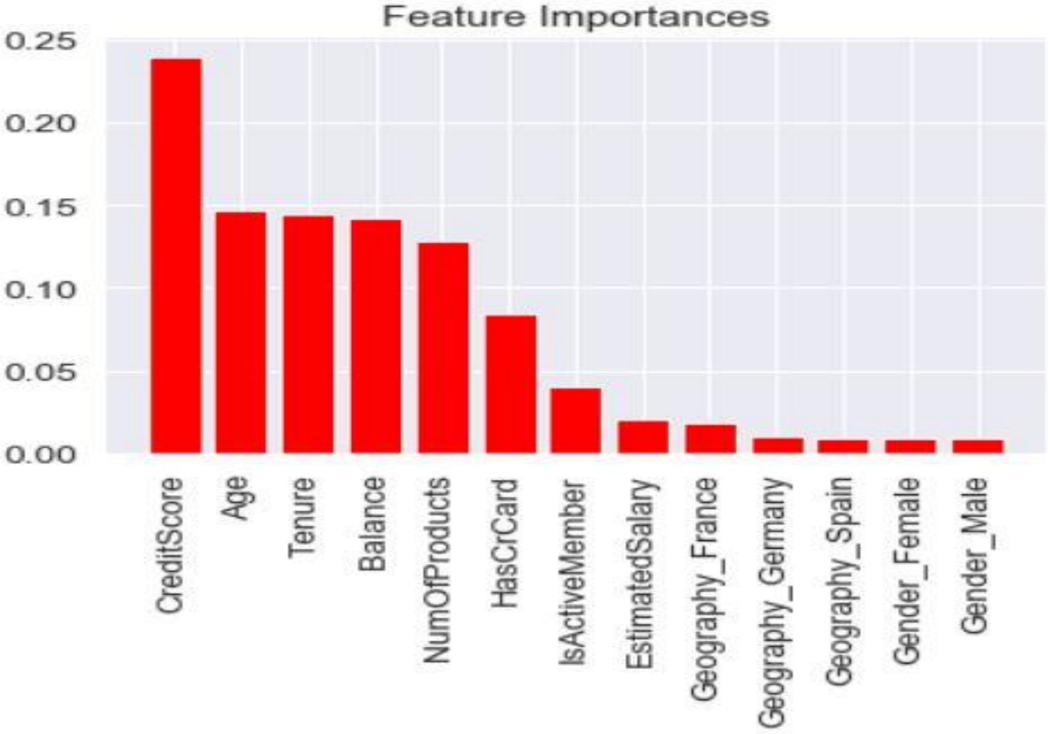
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$



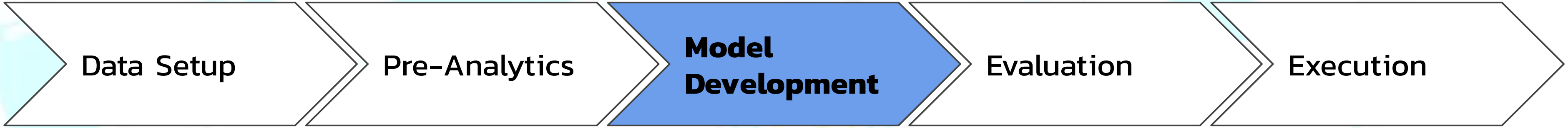
Heat map : Correlation Table



Feature Selection



- Random Forest Classification Algorithm
- 10,000 Decision Tree
- Gini Index


Data Setup
Pre-Analytics
Model Development
Evaluation
Execution

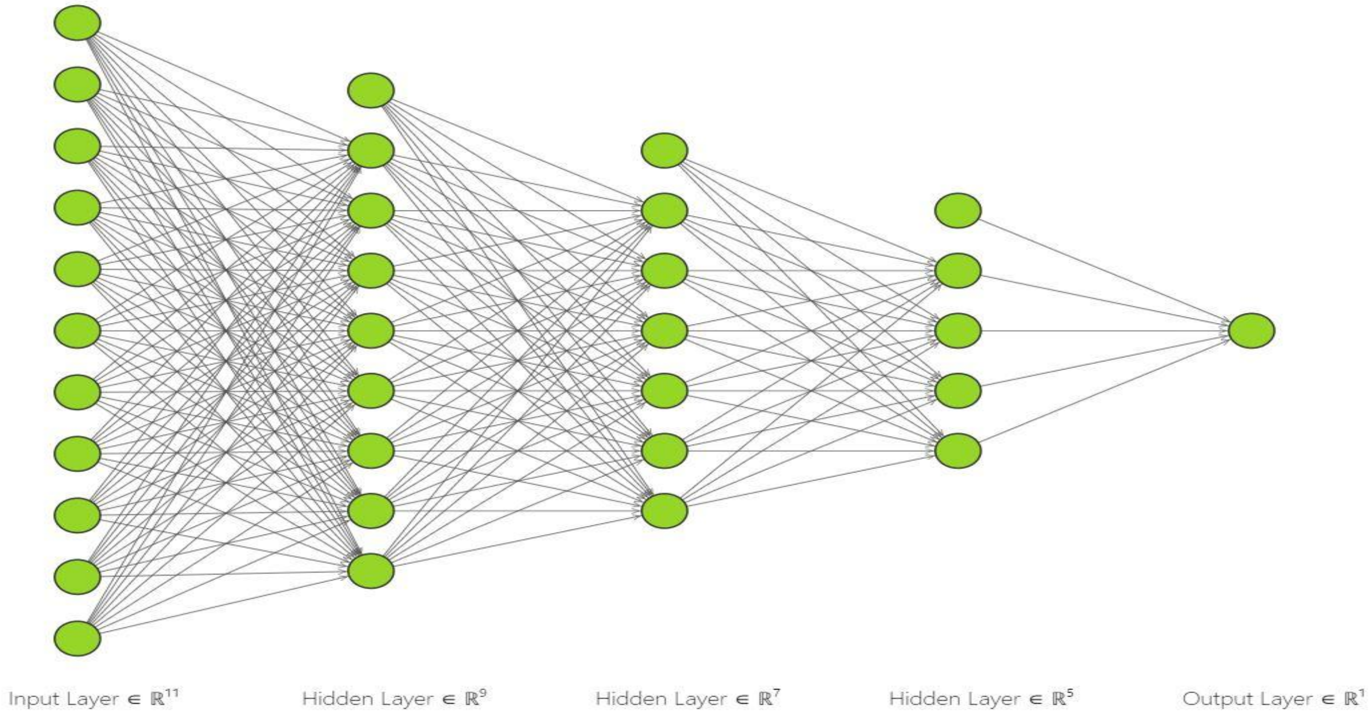
Why Neural Network?

- Neural Networks have the ability to learn by themselves and produce the output that is not limited to the input provided to them.
- These networks can learn from training data and apply them when a similar event arises, making them able to work through real-time events.
- Even if a neuron is not responding or a piece of information is missing, the network can detect the fault and still produce the output.
- They can perform multiple tasks in parallel without affecting the system performance.
- Corruption of one or more cells of ANN does not prevent it from generating output. This feature makes the networks fault-tolerant.

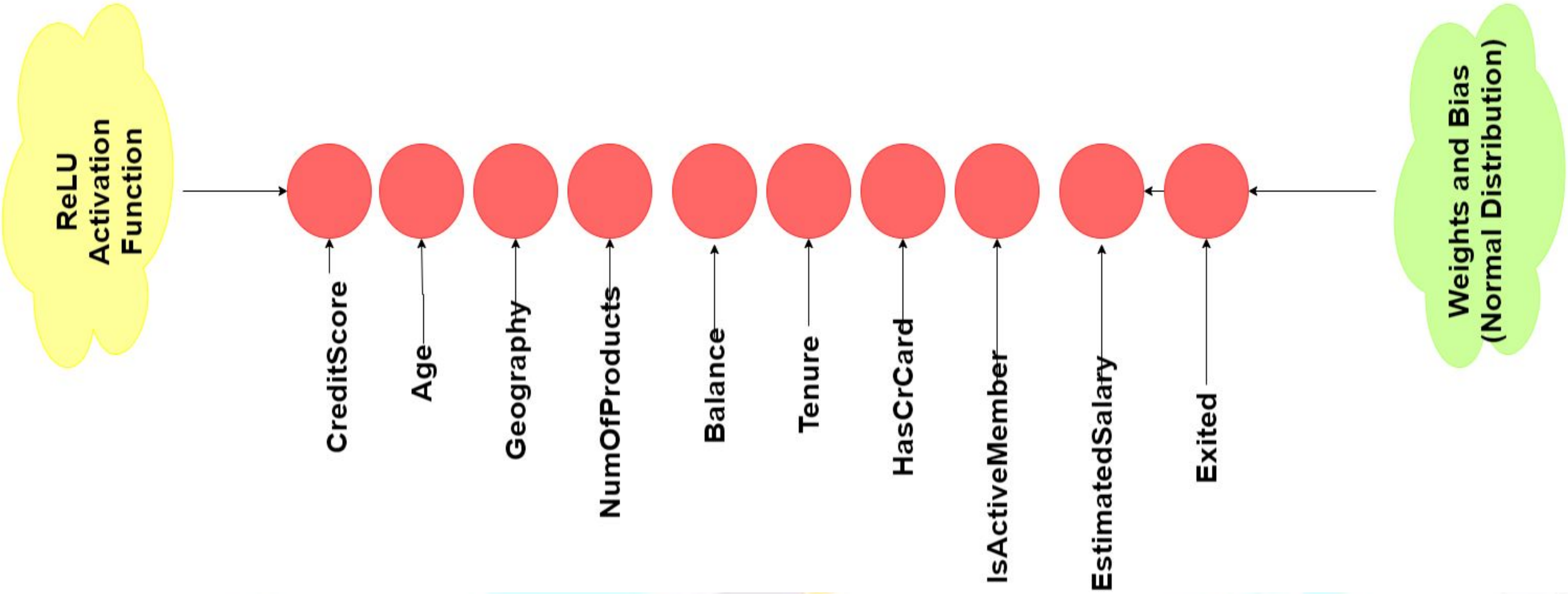
Neural Network Framework

Keras

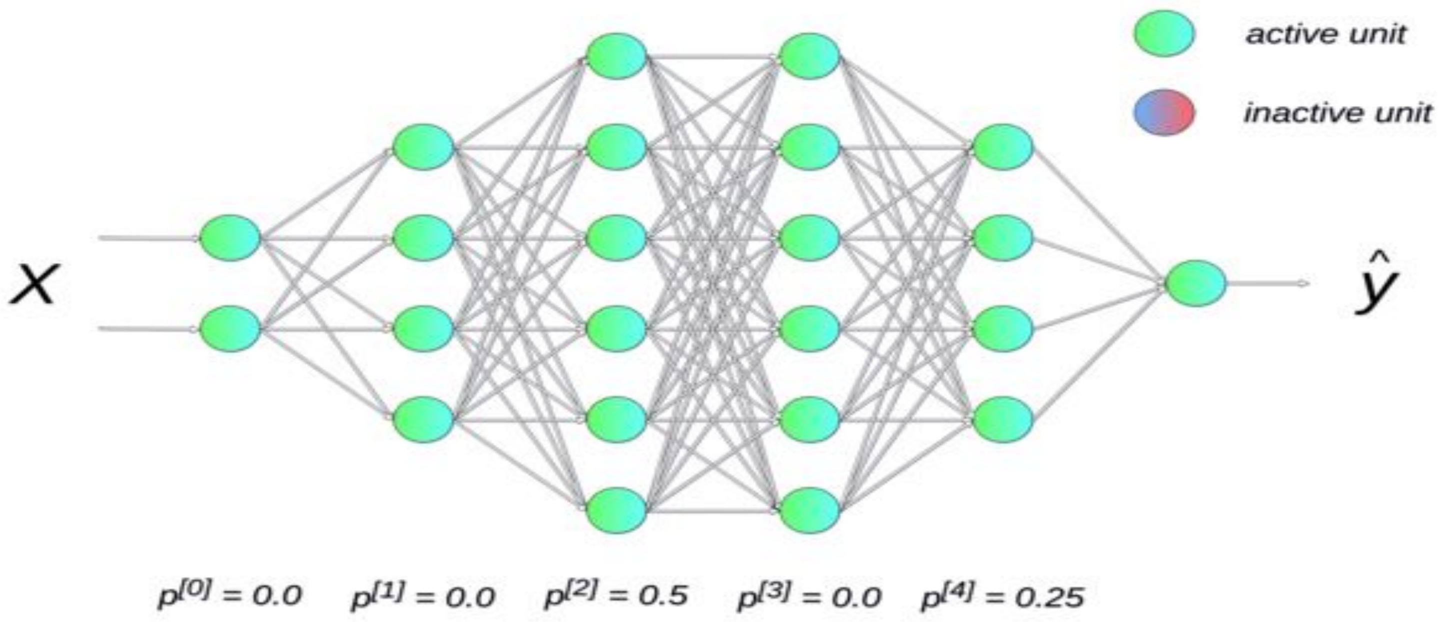
Neural Network Snapshot



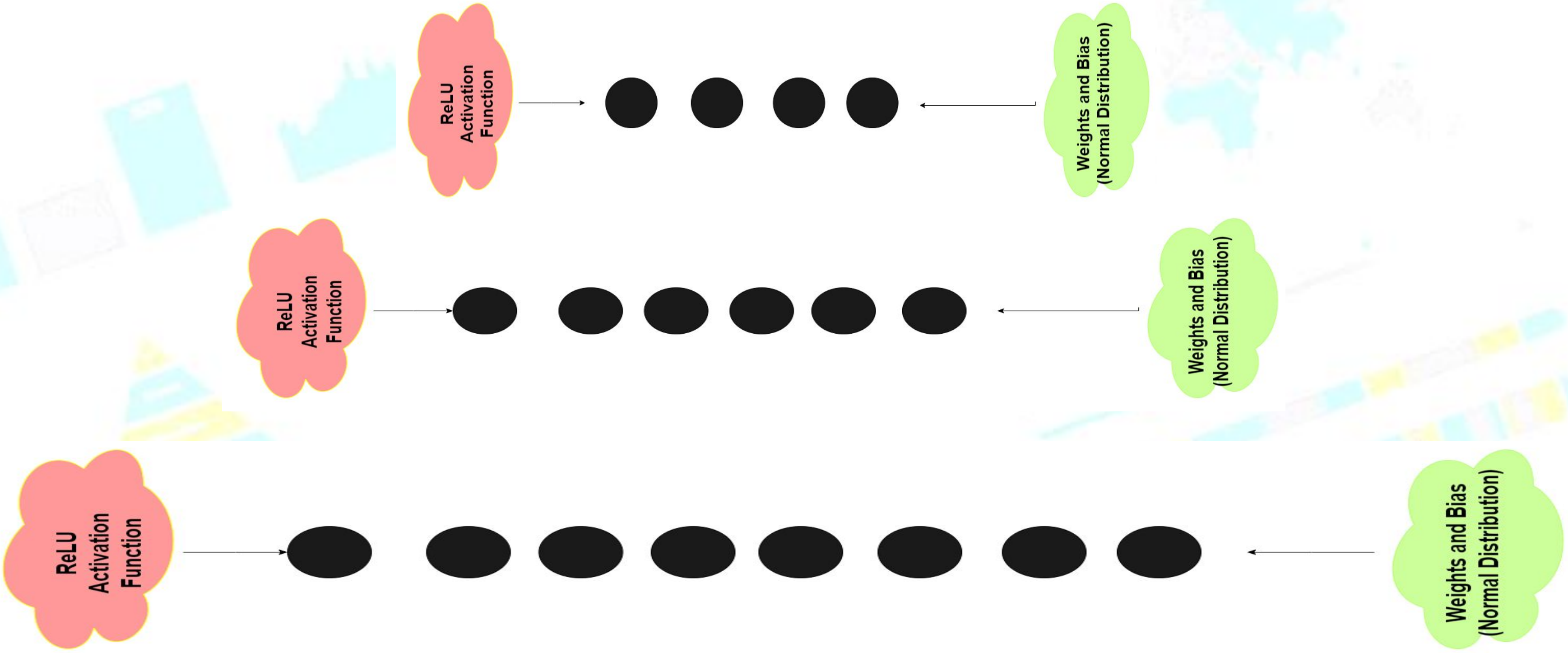
Input Layer


ReLU
Activation
Function
CreditScore
Age
Geography
NumOfProducts
Balance
Tenure
HasCrCard
IsActiveMember
EstimatedSalary
Exited
Weights and Bias
(Normal Distribution)

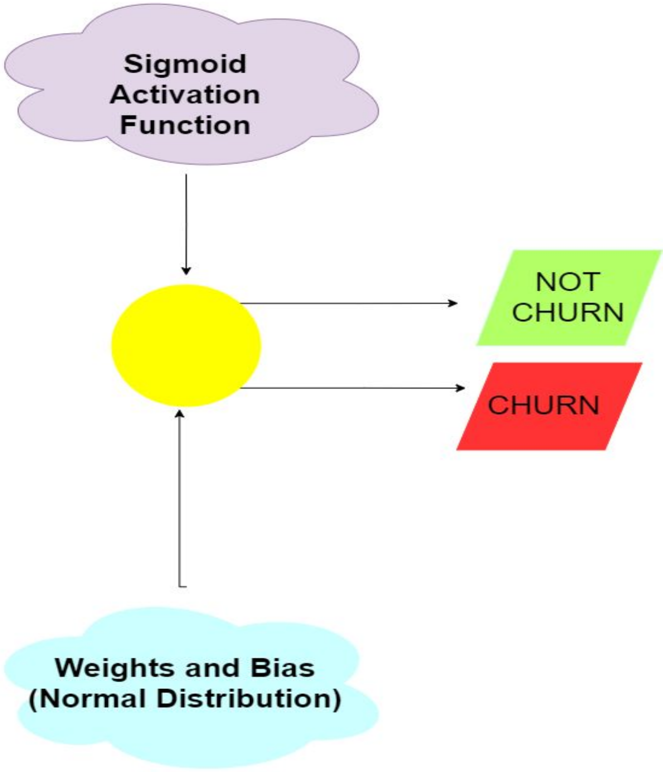
Dropout Layer

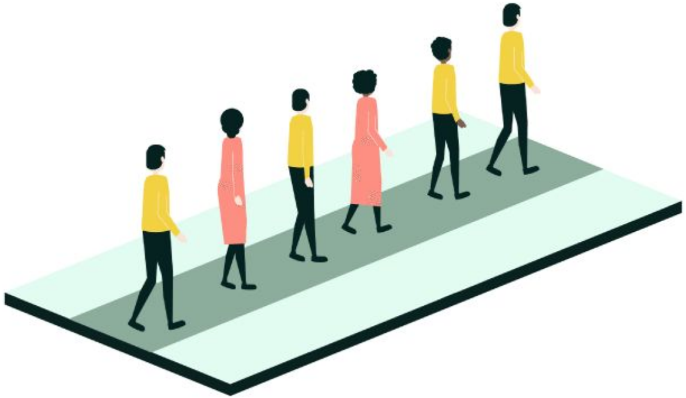


Hidden Layers

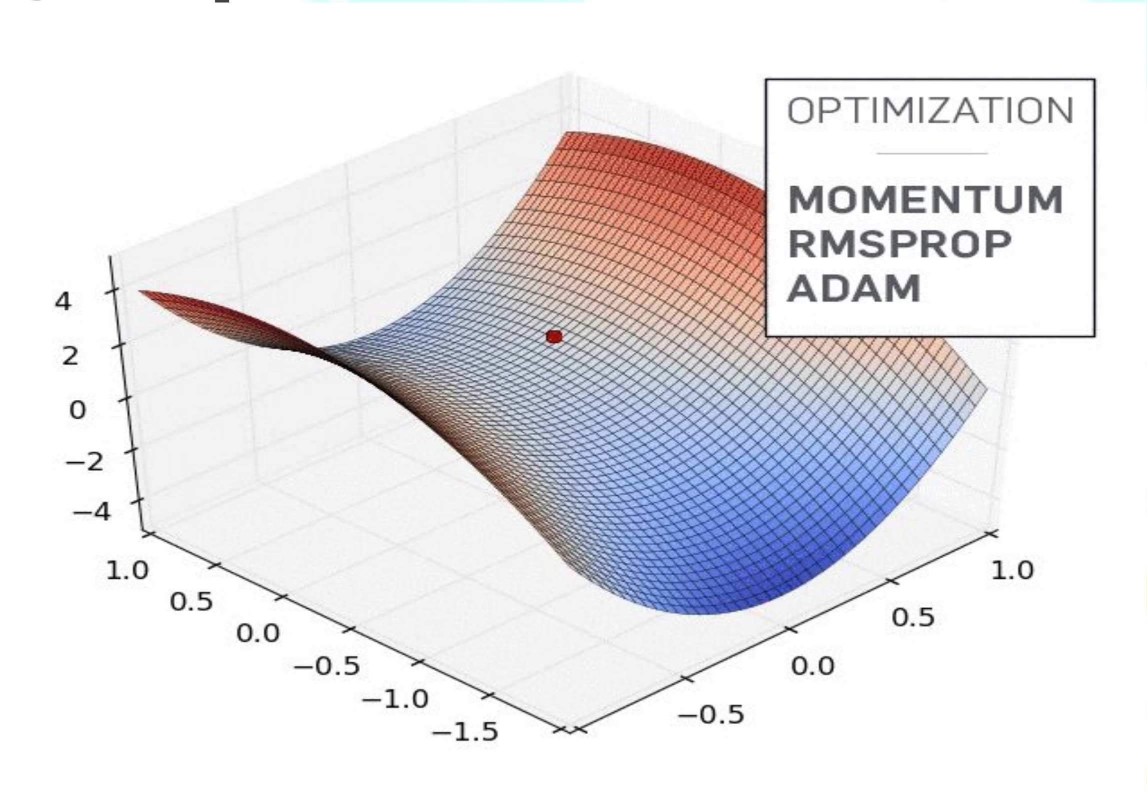


Output Layer



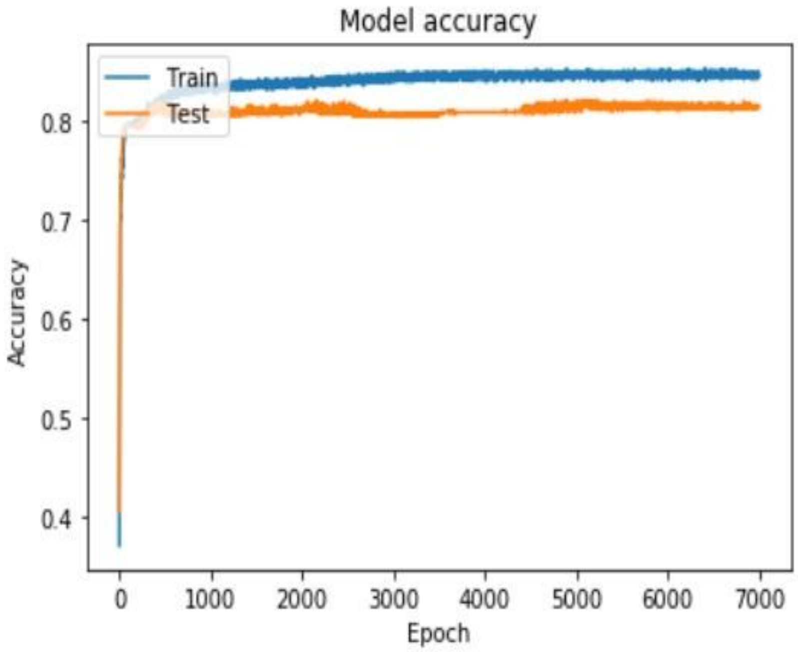


Weight Optimization



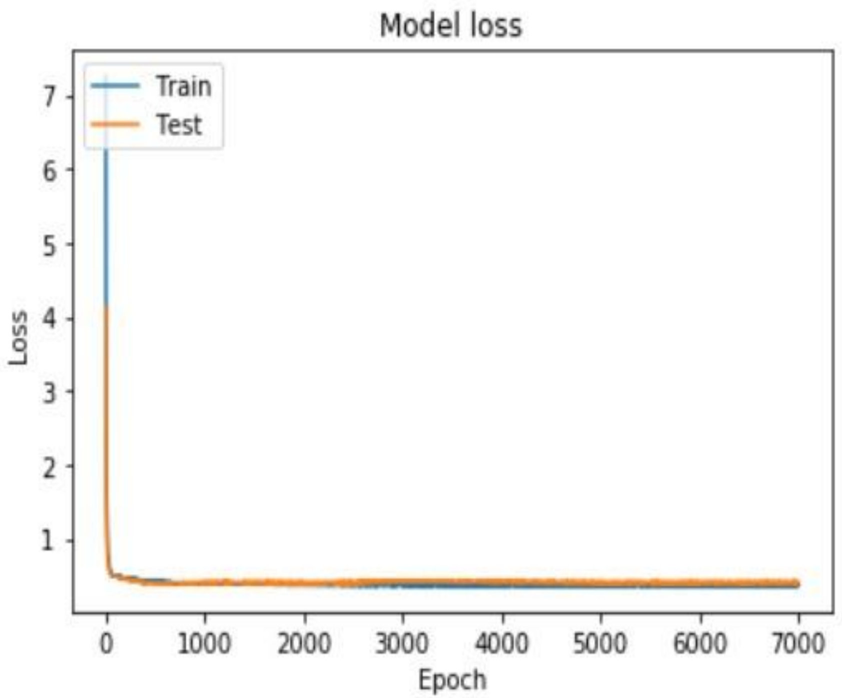
Data Setup → Pre-Analytics → Model Development → **Evaluation** → Execution

Accuracy Graph



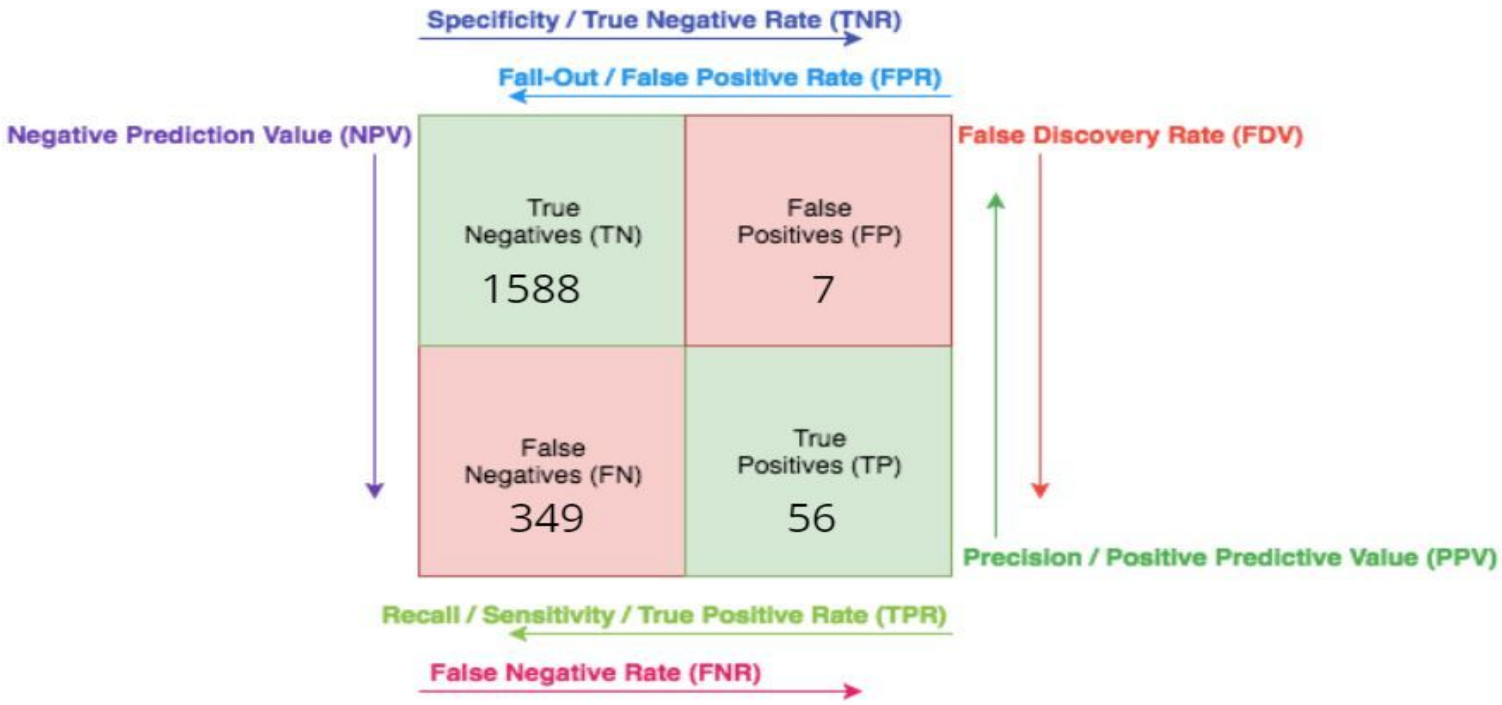
- The exponential increase in the accuracy of our model shows the the progress of training of NN

Loss Graph



- The representation of training and testing loss depicts that our model is trained correctly.
- Overfitting if: training loss << validation loss
- Underfitting if: training loss >> validation loss
- Just right if training loss ~ validation loss

Confusion Matrix



Confusion Matrix Measures

Measure	Value	Derivations
Sensitivity	0.82	TPR = TP / (TP + FN)
Specificity	0.89	SPC = TN / (FP + TN)
Precision	0.99	PPV = TP / (TP + FP)
Accuracy	0.83	ACC = (TP + TN) / (P + N)
F1 Score	0.89	F1 = 2TP / (2TP + FP + FN)

Data Setup → Pre-Analytics → Model Development → Evaluation → **Execution**

Project Details

https://github.com/rarpit1994/Churn-Prediction

Conclusion

- The Neural Network model was built based on certain important factors which resulted in accurate filtration of output layer.
- With increase in epochs , the accuracy of the model increased exponentially.
- The complex layers which worked in the backend was an important aspect because it formed a perfect base for the whole model to flourish exceedingly well.

References

- Brandusoiu, I., Toderean, G., & Beleiu, H. (2016). Methods for churn prediction in the prepaid mobile telecommunications industry. IEEE International Conference on Communications, 2016-August(April 2017), 97–100. https://doi.org/10.1109/ICComm.2016.7528311
- Sung, C., Higgins, C. Y., Zhang, B., & Choe, Y. (2017). Evaluating deep learning in chum prediction for everything-as-a-service in the cloud. Proceedings of the International Joint Conference on Neural Networks, 2017-May, 3664–3669. https://doi.org/10.1109/IJCNN.2017.7966317
- Bruce, 2011. (2013). 済無No Title No Title. Journal of Chemical Information and Modeling, 53(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004
- Günay, M. (2018). Makine Ö ğ renmesi Yöntemleri ile Kay ı p Mü ş teri Analizi Predictive Churn Analysis with Machine Learning Methods. 2018 26th Signal Processing and Communications Applications Conference (SIU)
- Chen, Y. B., Li, B. S., & Ge, X. Q. (2011). Study on predictive model of customer churn of mobile telecommunication company. Proceedings - 2011 4th International Conference on Business Intelligence and Financial Engineering, BIFE 2011, 114–117. https://doi.org/10.1109/BIFE.2011.112
- Kübra, Ş., & Güler, N. (2015). Bankac ı l ı kta Mü ş teri Terk Modeli Customer Churn Modelling in Banking.
- Martins Kummer, L. B., Cesar Nievola, J., & Paraiso, E. C. (2018). Applying Commitment to Churn and Remaining Players Lifetime Prediction. IEEE Conference on Computatonal Intelligence and Games, CIG, 2018-August. https://doi.org/10.1109/CIG.2018.8490443
- Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. Expert Systems with Applications, 36(10), 12547–12553. https://doi.org/10.1016/j.eswa.2009.05.032

Console Terminal Jobs

/cloud/project/

```
> print("Thank You !! :) ")
[1] "Thank You !! :) "
>
```