EE769: Building ML models from scratch

Project topic

Implement ML frameworks and their training algorithms efficiently from scratch such as:

- Random Forest
- SVM
- Neural Network

Overview and Background

1) Data extraction and preprocessing of data
2) Feature selection
3) Code development for the ML models
4) Training the model for the given data set
5) Evaluating the accuracy for the models
6) Comparing the accuracy with standard scikit-learn libraries in python.

Dataset

We used Customer Churn modelling Dataset. This dataset contain 10000 rows, with dependent variable being ‘Exited’. We have to predict whether a customer exited or not.

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

Preprocessing

During preprocessing we removed some of the columns which are having small correlation Coefficient

Converted Categorical values into numerical values

Normalise the dataset to make into similar range

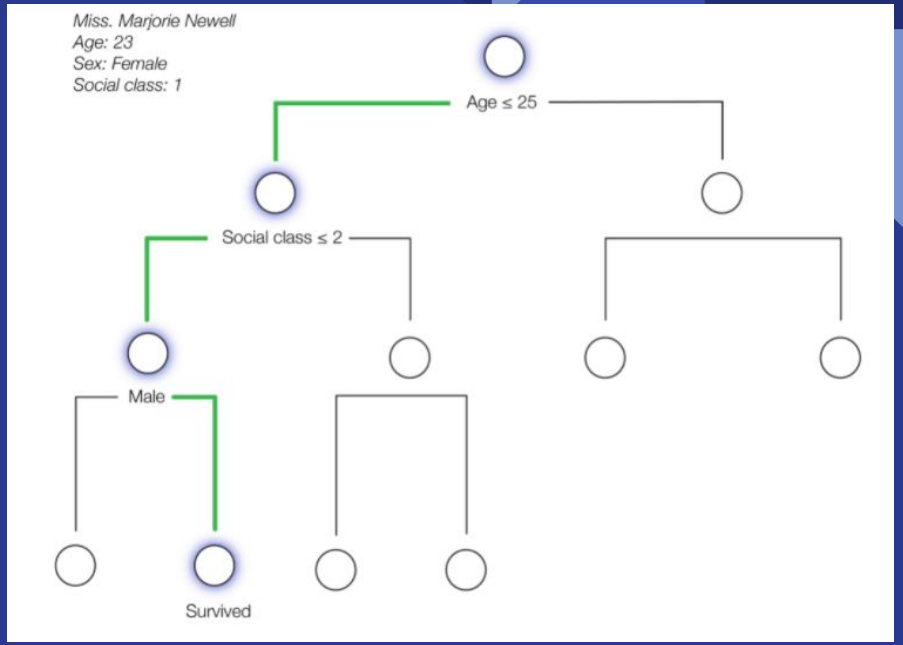
Implementation of models and their Accuracies:

Random Forest

- Random forests are known as ensemble learning methods used for classification and regression
- Random forests are essentially a collection of decision trees that are each fit on a subsample of the data.
- Random forests are also non-parametric and require little to no parameter tuning.
- For classification the terminal nodes of the decision tree output the class that is the mode while in the context of regression they'll output the mean prediction.

Random forest classifier

- Start from the root node and
  Move through the branches based on
  The question asked to reach the output.
- These individual trees are averaged to get the model.

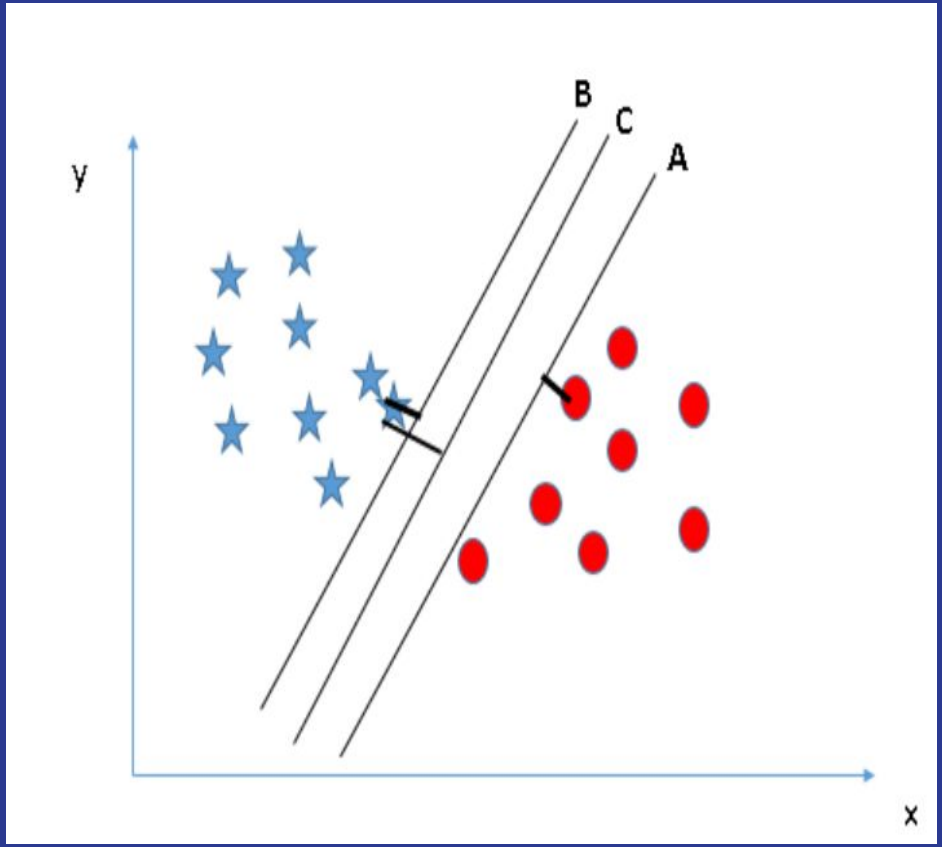


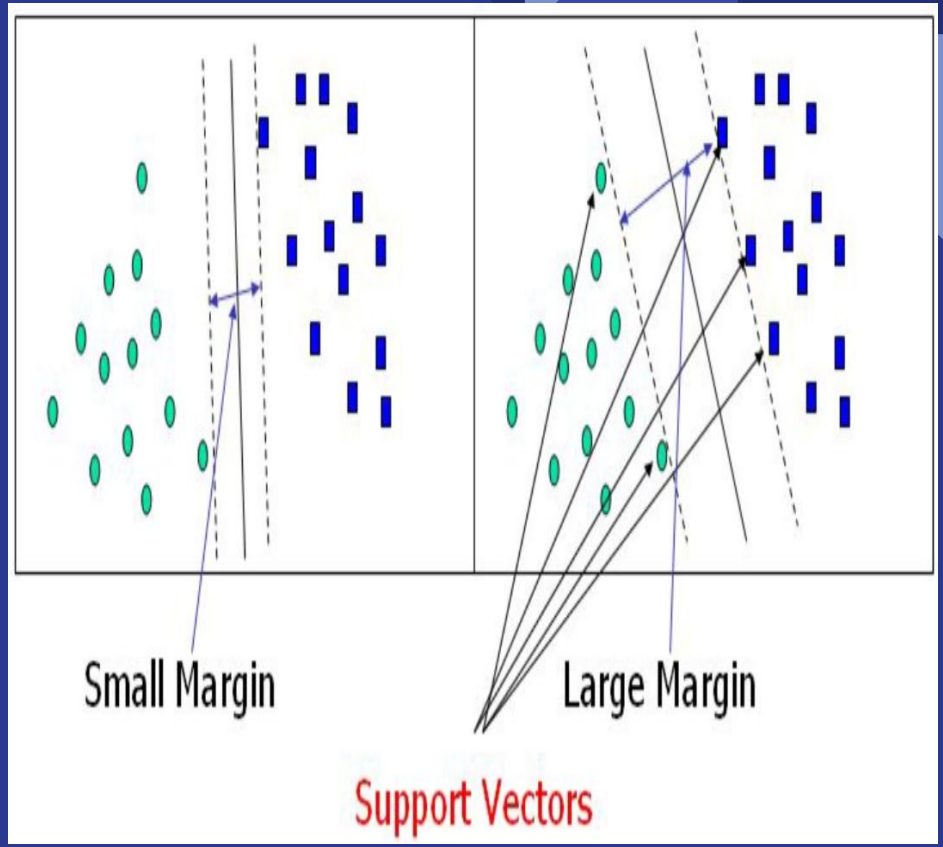
Support vector machines

- The SVM (Support Vector Machine) is a supervised machine learning algorithm typically used for binary classification problems.
- The algorithm finds a hyper-plane (or decision boundary) which should ideally have the following properties:
  * It creates separation between examples of two classes with a maximum margin.
  * The equation yields a positive value for +ve class and vice versa.

$$f(x) = sign\left(\mathbf{w}^* \cdot x + \mathbf{b}^*\right)$$

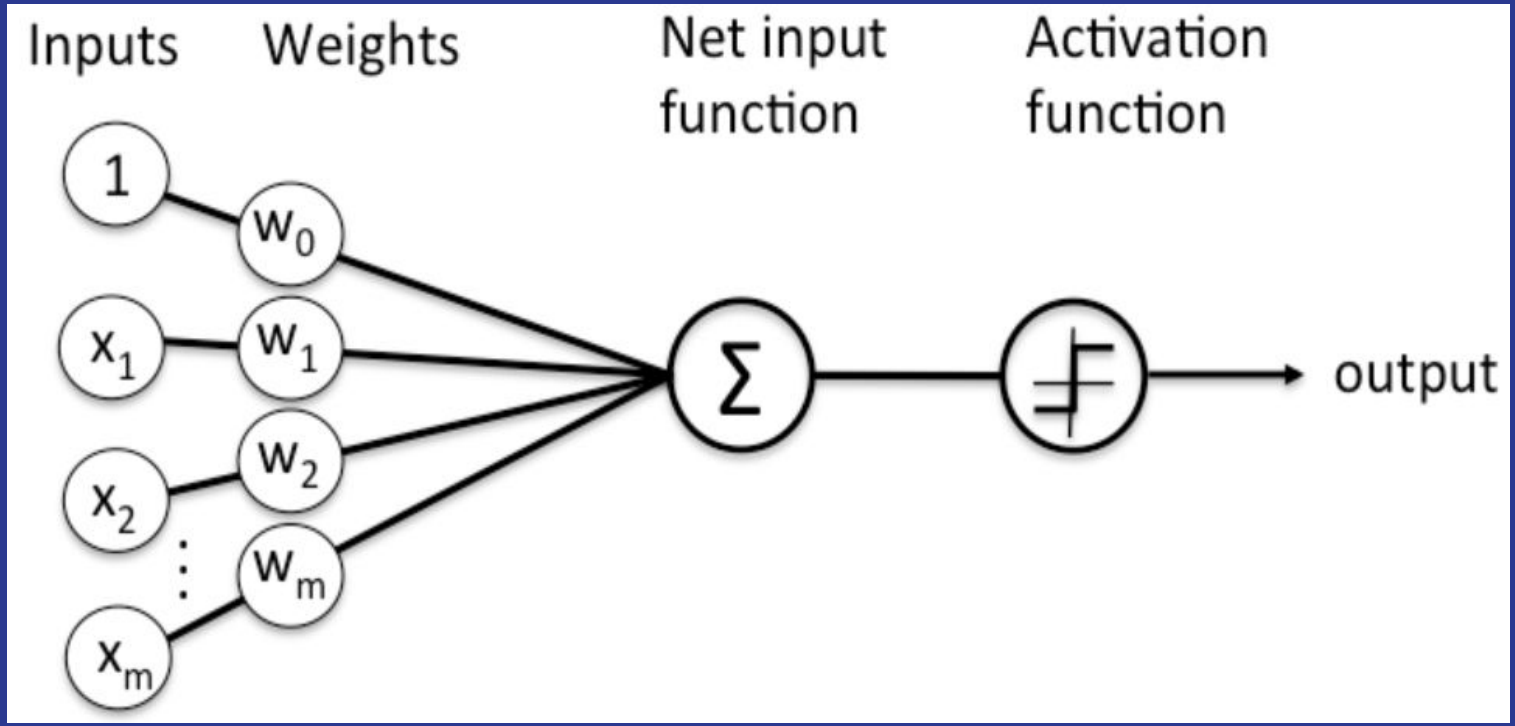
Support vector machines





Neural Network

- The two-layer neural network was implemented.
- Layers are made up of nodes. It is the place where calculation happens.

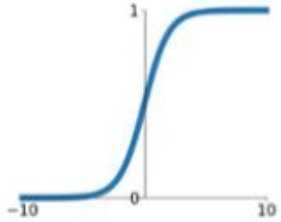


Neural Network

Activation Functions

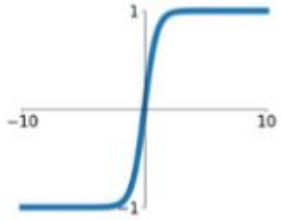
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



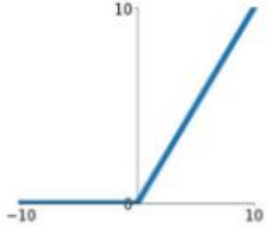
tanh

$$\tanh(x)$$



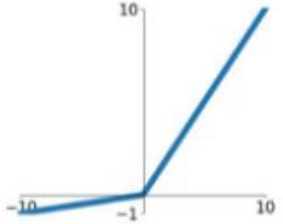
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

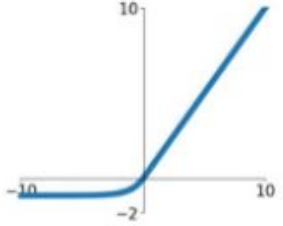


Maxout

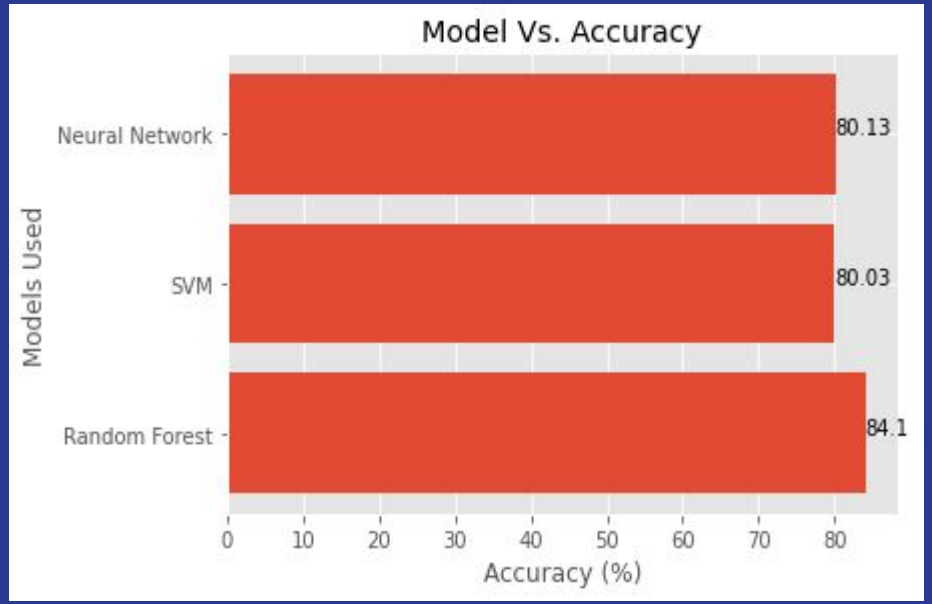
$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Best Classifier for Customer Churn Model Dataset



Best Classifier

Random Forest Classifier is the best classifier for Customer Churn Modelling Dataset.

Limitation and future work

- Lack of computational power.
- One could use CUDA which is parallel computing used for processing large blocks of data.