International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Performance Analysis of Predictive Models using Generic Datasets

Artika Singh, Manisha Jailia, Shubhangi Jain

Abstract: ***Today over 2.5 quintillion bytes of data is being created every single day where 753 crore people on this planet are creating 1.7mb of data each second. Most often than not, Researchers only scratch the surface when it comes to analyzing which algorithm will be best suited with their dataset and which one will give the highest efficiency. Sometimes, this analysis takes more computational time than the actual execution itself.***

Aim of this paper is to understand and solve this dilemma by applying different predictions models like Neural Networks, Regression and Decision Tree algorithms to different datasets where their performance was measured using ROC Index, Average Square Error and Misclassification Rate. A comparative analysis is done to show their best performance in different scopes and conditions. All data sets and results were compared and analyzed using SAS tool.

Keywords: Prediction Model, Decision Tree, Regression, Neural Network, ROC Index, Average Square Error, Misclassification Rate

I. INTRODUCTION

With this huge and increasing amount of data being generated in companies, the demand of Data analysis is surging rapidly. Companies have been employing the use of Data analysis in retail, healthcare, finance, manufacturing and transportation for a long time now. In this paper, we will be using data analysis to give a comparison between predictive algorithms on the basis of different data sets and their outputs.

A. Predictive Algorithm

1. Regression

Regression is a predictive modelling technique. It indicates the relationship between dependent and independent variable. [1] We will be employing the use of two regression techniques in this paper: Linear Regression and Logistic Regression.

1.1 Linear Regression

Linear Regression is a very simple model. It uses linear equation to indicate the relationship between the dependent and independent variable.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Artika Singh*, Perusing, PHD, Department, Computer Science and Engineering, Banasthali Vidhyapith.
Dr. Manisha Jailia, Associate Professor, Department, Computer Science Banasthali Vidhyapith.
Shubhangi Jain, Pursuing, B.Tech, Computer Engineering from MPSTME, NMIMS.

1.2 Logistic Regression

Logistic Regression is used to predict the value of categorical data like True/False. This means that the output variable is of the binary form. [1][2] It requires a large sample size to perform accurate results. It generates a finite number of outcomes. The main idea of Logistic Regression is to make the linear regression produce probabilities.

2. Decision Tree

Decision Tree is a type of supervised classification algorithm. [3] This algorithm is easy to implement and interpret as it generates the output in human-readable form. It extracts predictive information which can be easily understood by the human brain. This is one of the reliable classification techniques if reliable results are needed.

3. Neural Networks

Neural network is a series of algorithms that tries to imitate the way a human brain works. It is given a set of input data and it tries to figure out the relationship within these data sets. It is adaptable which means that it can adapt to changing input. [2][3] Some of the applications of Neural Networks is forecasting, market research for fraud detection and risk assessment.

B. Data Types

1. Interval

It is one of the highest levels of measurement in Steven's original system. It is quantitative in nature. This data type represents positions along continuous number lines rather than discrete number lines. One example of interval data type can be the Celsius temperature as the difference between each value is the same.

2. Categorical

Categorical data type, as the name suggests, has two or more categories. The data here is divided into categories and analysis is performed on these categories. This data type can be divided into two types: Nominal and Ordinal. [14] [15]

2.1 Nominal

In Nominal data type, no ordering is followed. These can be divided and grouped in categories but cannot be logically ranked or ordered. For example, Gender can be considered a categorical data type as it can be divided into two categories – Male and Female but it certainly cannot be logically ranked.

2.2 Ordinal

In Ordinal data type, a specific ordering is followed. This data type assigns observations into discrete categories. An example of ordinal data type from our daily lives can be the customer rating which represent the order such as very satisfied, satisfied, neutral, dissatisfied and disappointed.

3. Ratio

Having almost all the same properties as an interval data type, Ratio can be defined as a quantitative data. By using ratios, a logical rank, interval values plus the ability to calculate ratios as a true zero can be defined. Ratios provide a lot of possibilities for statistical analysis like performing arithmetic operations. It also helps in finding out the mean, median, mode and standard deviation.

C. Data Samples

1. Training Dataset

The input data which is fed to the algorithm in order to give an output is known as training data. The model learns from this data, tries to find pattern using this data, develop understanding and finally perform on its basis. [11] Thus, we can say that the training data is very important and should be reliable and accurate to give efficient results.

2. Validation Dataset

Once the model has been trained by the training set, the parameters and the final model are tuned with the help of validation set. The validation data set is the intermediate between the Training set and the Testing set. [11] [12] It is important as in this phase, the best model is selected and optimized.

3. Testing Dataset

This is the final phase which tests the performance of the model. This data set is only available during the testing phase. It is basically the evaluation of the final model while comparing it to other final models.

D. Analysis Statistics

1. ROC index

Receiver Operating Characteristics is the important performance measure stats for classification problem. It is a probabilistic curve which shows how good is our algorithm in separating different classes. [9] [12] Higher the values of ROC index better separating capability of the algorithm.

2. Average square error

The average square error basically calculates the distance between the regression line and a set of points. [11] The smaller this distance is the better. If the distance is smaller, we can deduce that the set of points is close to the line of best fit.

3. Misclassification error

This error if ignored, creates problems for analysis of data by giving biased decisions. This error should be minimized. [9] The parameters associated to this error should be well understood from studies in order to reduce this error.

II. METHODOLOGY

Three datasets, PVA97NK, ORGANIC and CREDIT has been taken having different target variables and processed through four prediction algorithms, Decision Tree, Probabilistic Decision Tree, Regression and Neural Network in SAS tool and compared the results on the basic of four statistics, ROC Curve, Average Square Error, Average Profit/Loss and Misclassification Error.

E. Datasets Used

A. PVA97NK [7][8]

This data has information about customer demographics, customer status, customer donations, charity promotions and target variable. Target A is to find the customer will do the charity or not, that is why it is a binary variable. Another Target D is also there for finding how much donation can be made. Here we are considering only Target A. In total we have 28 attributes in which we are using 27.

Name	Role	Level	Report
DemAge	Input	Interval	No
DemCluster	Input	Nominal	No
DemGender	Input	Nominal	No
DemHomeOwner	Input	Binary	No
DemMedHomeValue	Input	Interval	No
DemMedIncome	Input	Interval	No
DemPctVeterans	Input	Interval	No
GiftAvg36	Input	Interval	No
GiftAvgAll	Input	Interval	No
GiftAvgCard36	Input	Interval	No
GiftAvgLast	Input	Interval	No
GiftCnt36	Input	Interval	No
GiftCntAll	Input	Interval	No
GiftCntCard36	Input	Interval	No
GiftCntCardAll	Input	Interval	No
GiftTimeFirst	Input	Interval	No
GiftTimeLast	Input	Interval	No
ID	ID	Nominal	No
PromCnt12	Input	Interval	No
PromCnt36	Input	Interval	No
PromCntAll	Input	Interval	No
PromCntCard12	Input	Interval	No
PromCntCard36	Input	Interval	No
PromCntCardAll	Input	Interval	No
StatusCat96NK	Input	Nominal	No
StatusCatStarAll	Input	Binary	No
TargetB	Target	Binary	No
TargetD	Rejected	Interval	No

Fig.1 Dataset *PVA97NK*

B. ***Organic [7][8]*** In this data we have information about customer demographics, Organic product promotional value, time and class. There are two target variables are taken into account. TargetBuy is to find the customer will but the product or not, that is why it is a binary variable. TargetAmt is to find how much amount can be spending by customer on these products. In total we have 13 attributes in which we are using 12.

Name	Role	Level	Report
DemAffl	Input	Interval	No
DemAge	Input	Interval	No
DemCluster	Rejected	Nominal	No
DemClusterGroup	Input	Nominal	No
DemGender	Input	Nominal	No
DemReg	Input	Nominal	No
DemTVReg	Input	Nominal	No
ID	ID	Nominal	No
PromClass	Input	Nominal	No
PromSpend	Input	Interval	No
PromTime	Input	Interval	No
TargetAmt	Target	Interval	No
TargetBuy	Target	Binary	No

Fig.3 Dataset Credit

C. Credit[7][8]

Credit data has information about customer demographics, customer credit history and many other credit parameters. Target variable is to find the customer is eligible for credit or not, that is why it is a binary variable. In total we have 30 attributes and we are considering all of them.

Name	Role	Level	Report
BanruptcyInd	Input	Binary	No
CollectCnt	Input	Interval	No
DerogCnt	Input	Interval	No
ID	ID	Nominal	No
InqCnt06	Input	Interval	No
InqFinanceCnt24	Input	Interval	No
InqTimeLast	Input	Interval	No
TARGET	Target	Binary	No
TL50UtilCnt	Input	Interval	No
TL75UtilCnt	Input	Interval	No
TLBadCnt24	Input	Interval	No
TLBadDerogCnt	Input	Interval	No
TLBalHCPct	Input	Interval	No
TLCnt	Input	Interval	No
TLCnt03	Input	Interval	No
TLCnt12	Input	Interval	No
TLCnt24	Input	Interval	No
TLDel3060Cnt24	Input	Interval	No
TLDel60Cnt	Input	Interval	No
TLDel60Cnt24	Input	Interval	No
TLDel60CntAll	Input	Interval	No
TLDel90Cnt24	Input	Interval	No
TLMaxSum	Input	Interval	No
TLOpen24Pct	Input	Interval	No
TLOpenPct	Input	Interval	No
TLSatCnt	Input	Interval	No
TLSatPct	Input	Interval	No
TLSum	Input	Interval	No
TLTimeFirst	Input	Interval	No
TLTimeLast	Input	Interval	No

Fig.2 Dataset Organic

F. SAS Tools

SAS is a business intelligent tool which is licensed ware and provide may functionalities like Reporting, Analysis, Data Mining, Visualization, Predictive Modelling, Machine Learning and many more. It has a very interactive dashboard with excellent graphics which makes user very comfortable to work with.

I have many modules which works in the specific area like Base SAS, Enterprise Miner, Enterprise Guide, Business Visualization, Predictive Modelling and many more.

In industry it has high demand because of its technical capabilities and ease of learning. But the biggest drawback is its cost in the same time we have different options available in the market which provides almost same features and is a freeware.

G. Prediction Model

Model which can provide three basic tasks can be said as a predictive model:

- Able to convert its measures into prediction
- Capable of selecting useful inputs for the prediction
- End result should be an optimal model for prediction

All predictive model performs these steps to provide the end result. The comparative table of prediction algorithms which we use for analysis will give the better understanding and clarity of their workflow.

Essential Tasks	Decision Tree	Regression	Neutral Network
Conversions of measures into prediction	Calculate Logworh or entropy value	Linear Regression uses linear algebric equation Logistic Regression calculate Logit value with Logit equation	Uses linear regression equation with activation function having bias terms and weights
Selection of useful inputs	Through Information Gain or Gini Index	Through Sequential selection Forward Backward Stepwise	Can use Stepwise sequential selection but not used always due to its high computation
To find optimal model	Tree Pruning	Choose best model through sequence	Optimal model can be selected though "Stopped Training"

Table 1 - Comparison of Predictive Model

H. Fit Statistics Used

Measures are evaluated in three forms and for every form specific statistics are used to convert them into prediction. These forms could be as follows:

- *Decision* used for categorical values and provide end result like Yes or No. User need to select the best decision mapped in their case.
- *Ranking* gives the order or some sort of segregation which will help analyst to take the decision or predict the model outcome.
- *Estimation* provides a probabilistic value of each possible cases and accordingly user can rank it which leads to further in decision making.

Statistics used for above mention form of prediction can be easily viewed by self-explanatory table as follows:

Prediction Type	Fit Statistics	Optimal	Best Value
Decision	Misclassification Rate Average Profit/Loss	Value varies between 0 to 1, more close to 0 better, should be <0.4 close to 1 better / close to 0 good	Smallest Largest/ Smallest
Ranking	ROC Index	Value varies between 0 to 1, more close to 1 better, should be >0.7for strong model	Largest
Estimates	Average Square Error	Value varies between 0 to 1, more close to 0 better, should be <0.3	Smallest

Table 2 - Statistics Summary

III. MODEL COMPARISON

Each dataset was processed with all three algorithms and compared to find out the best predictive model for the data. These comparison is done on the basis of fit statistics results for validation data sample. The model provides best prediction for the given dataset is chosen as the best optimal model.

All these results are performed in SAS tool and the results and graphs are as follows: -

I. Dataset 1 (PVA97NK)

ROC index comparison of all the model decision tree, Probabilistic Decision tree regression and for training data set and validation data set which seems almost near to each other.

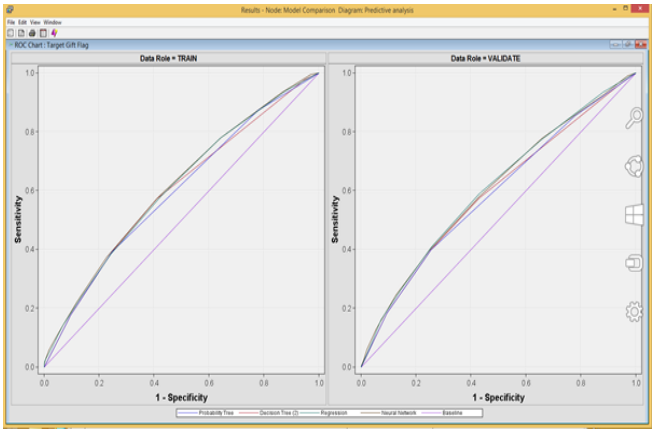

Fig.4 ROC Index comparison of Dataset *PVA97NK*

Statistical values will be give the better view of this comparison. In that considering average profit for validation data gave the better result. Best selected model is Neural network as it gives largest average profit for target variable in validation data.

```
*------------------------------------------------------------*
User:                  sasserver
Date:                  16 July 2019
Time:                  18:16:55
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*

Variable Summary

          Measurement    Frequency
 Role        Level         Count

TARGET      BINARY           1

Fit Statistics
Model Selection based on Valid: Average Profit for TargetB (_VAPROF_)

                                                Valid:      Train:
                                               Average     Average
Selected    Model                             Profit for  Squared
 Model      Node      Model Description        TargetB      Error

   Y        Neural    Neural Network           0.17328     0.23944
            Reg       Regression               0.17283     0.24088
            Tree2     Decision Tree (2)        0.15667     0.24194
            Tree3     Probability Tree         0.14293     0.24227

      Train:      Valid:
Misclassification  Average      Valid:
      Rate        Squared   Misclassification
                  Error        Rate

   0.49990        0.24141      0.50010
   0.49990        0.24139      0.50010
   0.49990        0.24338      0.50010
   0.49990        0.24327      0.50010
```

Fig.5 Best Model selection for Dataset *PVA97NK*

J. Dataset 2(Organic)

ROC index comparison of all the model decision tree, Probabilistic Decision tree regression and for training data set and validation data set in which except regression others seems almost near to each other.

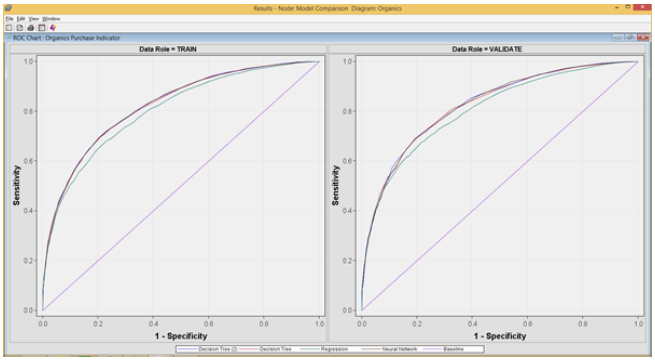

Fig.6 ROC Index comparison of Dataset *Organic*

Statistical values will be give the better view of this comparison. In that considering misclassification rate for validation data gave the better result. Best selected model is Decision tree as it gives smallest misclassification rate for target variable in validation data.

```
*------------------------------------------------------------*
User:                  sasserver
Date:                  17 July 2019
Time:                  13:33:29
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*

Variable Summary

          Measurement    Frequency
 Role        Level         Count

TARGET      BINARY           1

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

                                                              Train:
Valid:                                            Valid:      Average
Train:      Average
Selected  Model                                 Misclassification  Squared
Misclassification  Squared
 Model    Node    Model Description                 Rate          Error
Rate        Error

   Y      Tree    Decision Tree                  0.18531       0.13286
0.18512       0.13277
          Neural  Neural Network                 0.18585       0.13325
0.18332       0.13321
          Reg     Regression                     0.18612       0.13955
0.18781       0.13820
          Tree2   Decision Tree (2)              0.18765       0.13301
0.18476       0.13266
```

Fig.7 Best Model selection for Dataset *Organic*

K. Dataset 3(Credit)

ROC index comparison of all the model decision tree, Probabilistic Decision tree regression and for training data set and validation data set in which all algorithms have distinct values and curves. In both the cases Neural Network gives highest ROC index and Decision Tree gave the lowest.

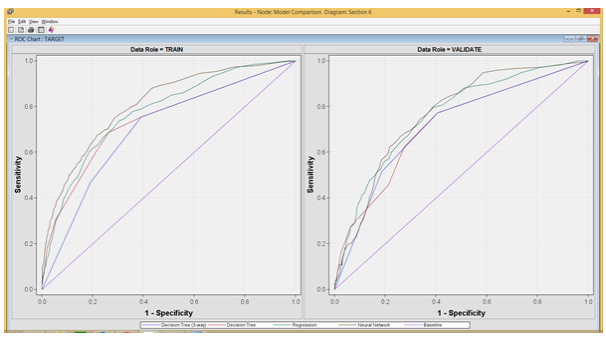

Fig.8 ROC Index comparison of Dataset *Credit*

Statistical values will be give the better view of this comparison. In that considering Average square error for validation data gave the better result. Best selected model is Regression as it gives smallest Average square error for target variable in validation data.

```
*-------------------------------------------------------------*
User:                  sasserver
Date:                  22 July 2019
Time:                  23:30:01
*-------------------------------------------------------------*
* Training Output
*-------------------------------------------------------------*

Variable Summary

          Measurement    Frequency
 Role        Level         Count

TARGET      BINARY           1
|
Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

                                          Valid:   Train:
                                         Average  Average
Train:              Valid:
Selected  Model                          Squared  Squared
Misclassification  Misclassification
 Model     Node    Model Description      Error    Error
Rate                Rate

   Y       Reg     Regression            0.12221  0.11756
0.16051             0.17333
          Tree    Decision Tree         0.12471  0.11729
0.15385             0.16095
          Neural  Neural Network        0.12546  0.10951
0.14513             0.17238
          Tree2   Decision Tree (3-way) 0.12584  0.12747
0.16667             0.16667
```

Fig.9 Best Model selection for Dataset *Credit*

IV. RESULT ANALYSIS

ROC curve of Data set 1 was not much strong and neither clearly separating those algorithms, which means segregation of classes are not that much proper. In that case other stats were taken into account for decision making. Average profit gives provides clear result as Neural Network was the best for prediction.

In Data set 2 perfect ROC curve were created except one rest of algorithm were merging but this shows classification can be done properly between true and false results. Misclassification rate gives good result and Decision Tree was select as best fit model.

Data set 3 ROC curve provided proper separation of all the algorithms but only few algorithms gave standard curve. In validation data set Regression and Neural Network are providing close results to avoid ambiguity another stat Average Square Error taken into consideration which gives Regression as a best predictive model.

There are other findings gathered from the above comparison of results which are as follows: -

	Decision Tree	Regression	Neural Network
Data Type (Target variable)	Good with non-numeric values	Non-numeric values can be a big trouble	Both numeric and non-numeric handled well
Missing Values	Can handle well	Can't handle	Can't handle
Extreme Values	Does not affect much	Very much sensitive, Transformation required	Activation function take care of extremes
Large no. of input variable	Easily taken care of through pruning	Could increase complexity but sequence selection can manage	Increases complexity and require very high computation
Independent variable	Better than regression	Not good with	Handles pretty well

Table 3 - Findings from comparative results

V. CONCLUSION

This paper discusses general (theoretical) and as well as practical (tool based) comparison of different algorithms with distinct data sets and provide some insights on which kind of predictive models are best suited for which kind of dataset. Also thrown some light on well suited statistics for decision making in different scenarios.

Target variable's data type plays an important role in algorithm selection. Even the target variable data type is same for different data sets still optimal model could be different from them. Because it also depends on which kind of statistics were used for model selection.

When Average Profit/Loss is taken into account for result analysis then other statistics will never get preference over that whichever model give best performance in that will be selected.

In general performance accuracy of Neural Network is better than Regression but as complexity and computation power of Neural network is very high that is why Regression is preferred in some cases.

REFERENCES

1. K Prasanna Jyothi, Dr R SivaRanjani, Dr Tusar Kanti Mishra, S Ranjan Mishra4 ,"A Study of Classification Techniques of Data Mining Techniques in Health Related Research",International Journal of Innovative Research in Computer and Communication Engineering,Vol. 5, Issue 7, July 2017.
2. Beg¨um Çı˘gs¸ar and Deniz ¨Unal,"Comparison of Data mining classification algorithms determining the default risk", Hindawi,Scientific Programming,Volume 2019,Article ID: 8706505.
3. Sagar S. Nikam," A comparative study of classification techniques in data mining algorithms", Orient. J. Comp. Sci. & Technol., Vol. **8**(1), 13-19 (2015).
4. Rohit Arora, Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA",International Journal of Computer Applications (0975 – 8887) ,Volume 54– No.13, September 2012.
5. Hetal Bhavsar, Amit Ganatra,"A Comparative Study of Training Algorithms for supervised machine "International Journal of Soft Computing and Engineering (IJSCE),ISSN: 2231-2307, Volume-2, Issue-4, September 2012.
6. Mohamed Aly,"Survey on Multiclass Classification Methods",unpublished.
7. SAS Institute Inc., Getting Started with SAS® Enterprise Miner™ 14.1, Cary, North Carolina 27513-2414, Copyright © 2015
8. SAS Institute Inc.,Applied analytics using SAS® Enterprise Miner™ Course notes, Cary, North Carolina, Copyright © 2010.
9. Jiawei Han,Micheline Kamber, Jian Pei,Data mining concepts and techniques,3rd edition, Morgan Kaufmann,2011
10. Delveen Luqman Abd Al.Nabi, Shereen Shukri Ahmed, "Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)",(ISSN 2222-2863),Vol.4, No.8, 2013
11. Z. Haiyang, "A Short Introduction to Data Mining and Its Applications", IEEE, 2011
12. K. P. Soman, Insight into Data Miining Theory and Practice, New Delhi: PHI, 2006.
13. XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, ―Top 10 algorithms in data mining,‖ *Knowledge Information system*, vol. 14, pp. 1-37, 2008.
14. SMs. Aparna Raj, Mrs. Bincy, Mrs. T.Mathu "Survey on Common Data Mining Classification Techniques", International Journal of Wisdom Based Computing, Vol. 2(1), April 2012
15. Raj Kumar, Dr. Rajesh Verma," Classification Algorithms for Data Mining P: A Survey" IJIET Vol. 1 Issue August 2012, ISSN: 2319 – 1058.

16. S.Archana, Dr. K.Elangovan, “Survey of classifications techniques in Data Mining”, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014
17. Vivek Agarwal, Saket Thakare, Akshay Jaiswal, "Survey on Classification Techniques for Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 132 – No.4, December2015.

AUTHORS PROFILE

Artika Singh, perusing PHD from Banasthali vidhyapith, did M.Tech in computer science and engineering, currently working in Data Science and Data Analytics filed. Have 08 publications in this domain which are published in IEEE, Springer and others.

Dr. Manisha Jailia, PHD in computer science department and working as Associate Professor in Banasthali Vidhyapith. Her research areas are Data Mining, Distributed Databases, Web technologies, Big Data and Cloud Computing.

Shubhangi Jain, is currently pursuing B.Tech in Computer Engineering from MPSTME, NMIMS. She has previously worked on two research papers which have been presented and will be published in IEEE Xplore by May 2020. She has done two internships related to her course.