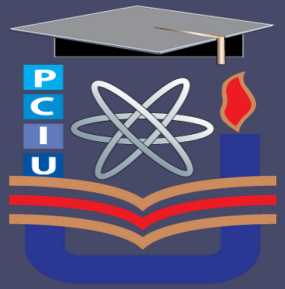


Bank Customer Segmentation Using Machine and Deep Learning Methods

Proposed By:

Name: Niloy Mitra
ID:CSE 01806713
Batch: CSE 18-D-B

Supervised By:

Mrs. Manoara Begum
Assistant Professor
Dept. Of CSE

Contents

- ❑ Introduction
- ❑ Motivation
- ❑ Objectives
- ❑ Literature Review
- ❑ Methodology
- ❑ System Tools
- ❑ Data Collection
- ❑ Data Preprocessing
- ❑ Classification Algorithm
- ❑ Result & Discussion
- ❑ Comparison With Previous Work
- ❑ Conclusion and Future Work
- ❑ References

Introduction

- Customer Segmentation is whether predicting the customers are actually a customer of that bank.
- By providing their customers with high quality banking services and high technology innovative products.
- Banks are continuously trying to retain their existing customer and attract new customers by providing better quality services
- Machine learning and deep learning model has been used.

Motivation

- Scopes for Improving Work Through Increasing accuracy and enhancing the Dataset.
- This thesis investigates the Customer classification problem for prediction, employing supervised approaches to determine the overall semantic of customer prediction by classifying them as Yes or No.
- A model could be assist possible clients with settling on an informed decision on their get services from the banks and banks to improve their services.
- Fewer deep learning model used on Bank Customer dataset.

Objective

- My research aims to do this by conducting Bank Customer prediction or not as Yes or No.
- To examine the relationship between Customer and Bank.
- It can be beneficial to the bank because bank can easily understand who are their customers and who are not their customers.
- To find the highly effective model with Bank Customer Prediction dataset.

Literature Review

NO	Paper Title	Author	Publisher and Year	Findings
01	Predicting Satisfaction of Online Banking System Using ML	Mohammad Monirul Islam, Refath Ara Hossain , Nazmun Nessa Moon and Fernaz Narin Nur	IEEE, 2021	**Dataset:** 3526 Records. **Algorithm:**RF,LR,DT,ANN,KNN. **Accuracy:**Highest accuracy was 96%. **Limitations:**Deep Learning Can not be Used and Dataset Size is too low.
02	Prediction of Potential Bank Customers: Application on Data Mining	Muhammet Sinan Başarslan İrem Düzdar Argun	Springer, 2020	**Dataset:** 45211 Records. **Model:**NB,LR,RF,KNN. **Accuracy:** Highest accuracy was 93%. **Limitations:** They did not show the graph well and Use only 4 Algorithm.

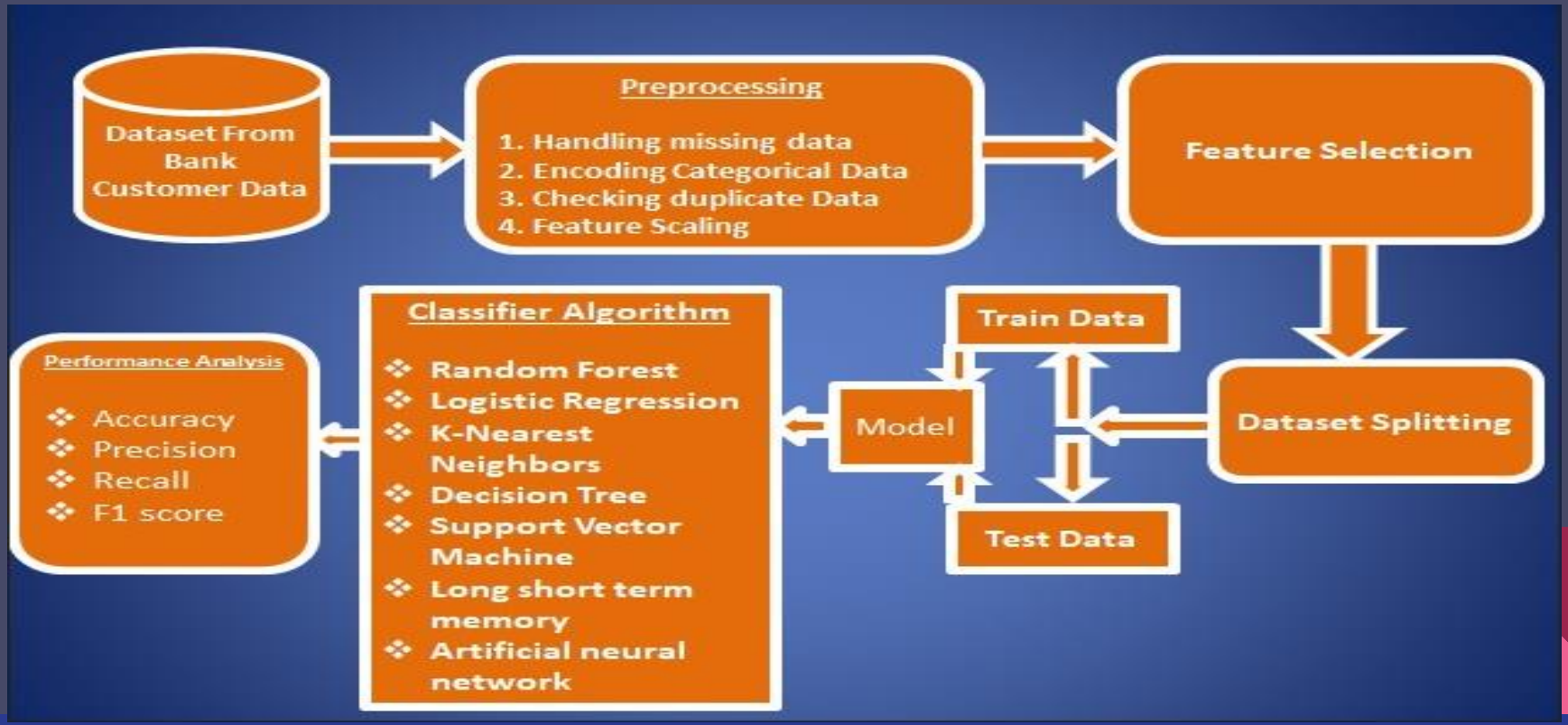
Literature Review

NO	Paper Title	Author	Publisher and Year	Findings
03	Predicting Bank Depositor's Behavior	Fereshteh Safarkhani, Sérgio Moro	Multidisciplinary Digital Publishing Institute (MDPI), 2021	**Dataset:** Size 4119 **Algorithm:** DT **Accuracy:** 90.87%(DT) **Limitations:** They Use only one algorithm and dataset can be imbalenced.
04	Predicting the Future Transaction from Large and Imbalanced Banking Dataset	Sadaf ilyas Sultan zia Umair Muneer Sukumar Letchmunan	ResearchGate, 2020	**Dataset:** KAGGLE Dataset **Algorithm:**LR,DT,RF. **Accuracy:** Highest accuracy was 91%. **Limitations**: Deep learning can not be used and They described dataset can be imbalenced.AUC Result will less.

Literature Review

NO	Paper Title	Author	Publisher and Year	Findings
05	Machine-Learning Techniques for Customer Retention: A Comparative Study	Sahar F. Sabbeh	ResearchGate, 2018	**Dataset:** 3333 Records **Algorithm:**RF,SVM,KNN,NB,LR. **Accuracy:** Highest accuracy was 92.85%. **Limitations:** Deep learning can not be used.The Dataset is Small.

Methodology



System Tools

- Python
- Pandas
- NumPy
- Matplotlib
- Scikit-learn
- Google Colab

Data Collection

I Collect the data has been used in this study is taken from Kaggle. We have collected the Data from Kaggle is Bank Customer Data. There are 17 properties and 42639 customer records in the bank data set. Our Dataset also had numerical and Categorical Data. This dataset has been labeled with two classes, which are Yes Or No.

Data Preprocessing

❑ Data Before Encoding

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	Customer
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

❑ Data After Encoding

	age	balance	day	duration	campaign	pdays	previous	Customer	job_blue-collar	job_entrepreneur	...	month_may	month_nov	month_oct	month_sep	loan_yes	contact_telephone	contact_unknown	poutc
0	58	2143	5	261	1	-1	0	no	0	0	...	1	0	0	0	0	0	1	
1	44	29	5	151	1	-1	0	no	0	0	...	1	0	0	0	0	0	1	
2	33	2	5	76	1	-1	0	no	0	1	...	1	0	0	0	1	0	1	
3	47	1506	5	92	1	-1	0	no	1	0	...	1	0	0	0	0	0	1	
4	33	1	5	198	1	-1	0	no	0	0	...	1	0	0	0	0	0	1	

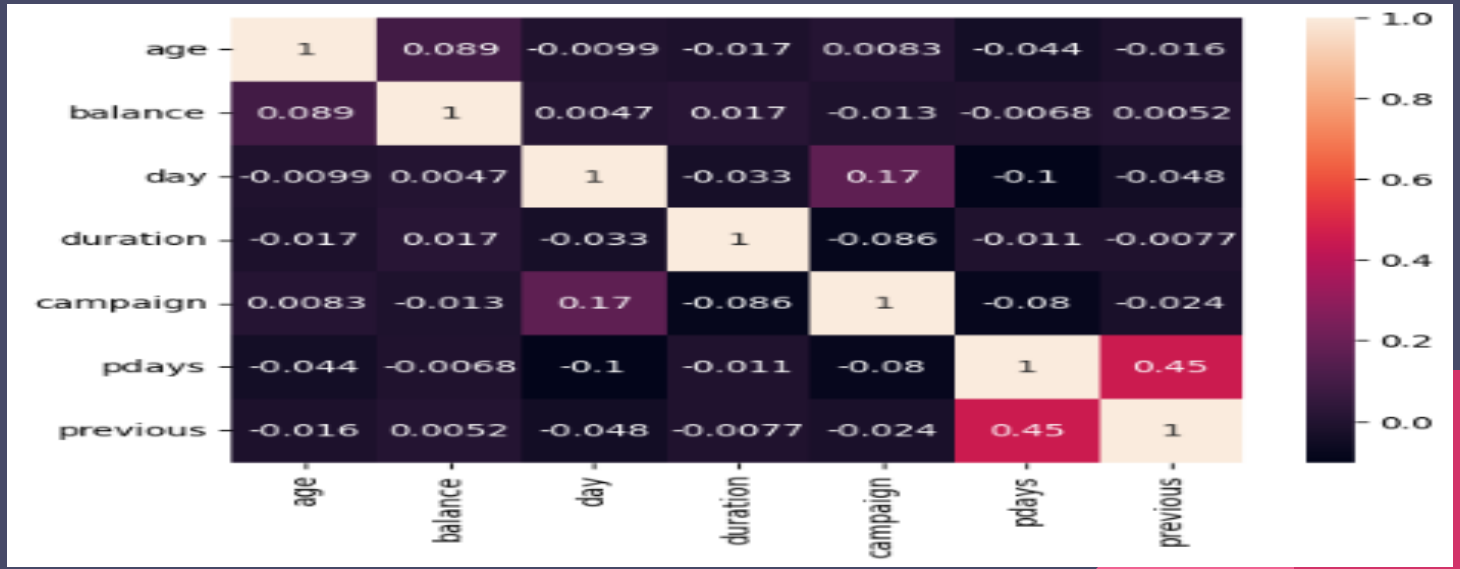
Feature Scaling

- Feature Scaling shrinks the data within the given range, usually of 0 to 1.
- It transforms data by scaling features to a given range.

	age	balance	day	duration	campaign	pdays	previous	job_blue-collar	job_entrepreneur	job_housemaid	...	month_may	month_nov	month_oct	month_sep	loan_yes	contact_telephone	contact_unknown
0	0.519481	0.092259	0.133333	0.053070	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	1.0
1	0.337662	0.073067	0.133333	0.030704	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	1.0
2	0.194805	0.072822	0.133333	0.015453	0.0	0.0	0.0	0.0	1.0	0.0	...	1.0	0.0	0.0	0.0	1.0	0.0	1.0
3	0.376623	0.086476	0.133333	0.018707	0.0	0.0	0.0	1.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	1.0
4	0.194805	0.072812	0.133333	0.040260	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	1.0

Feature Selection

- Feature Selection is the process used to select the input variables that are most important to Machine Learning task.
- Here we have used Correlation Matrix as Feature Selection.

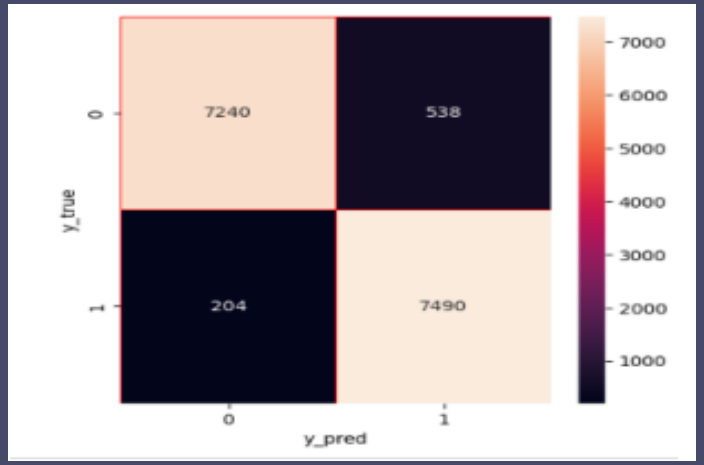


Classification Algorithm

- **Machine Learning Model**
  - Random forest (RF)
  - Logistic Regression (LR)
  - K-Nearest Neighbor(KNN)
  - Decision Tree (DT)
  - Support Vector Machine(SVM)
- **Deep learning Model**
  - Artificial neural network(ANN)
  - Long Short-Term Memory(LSTM)

Result & Discussion

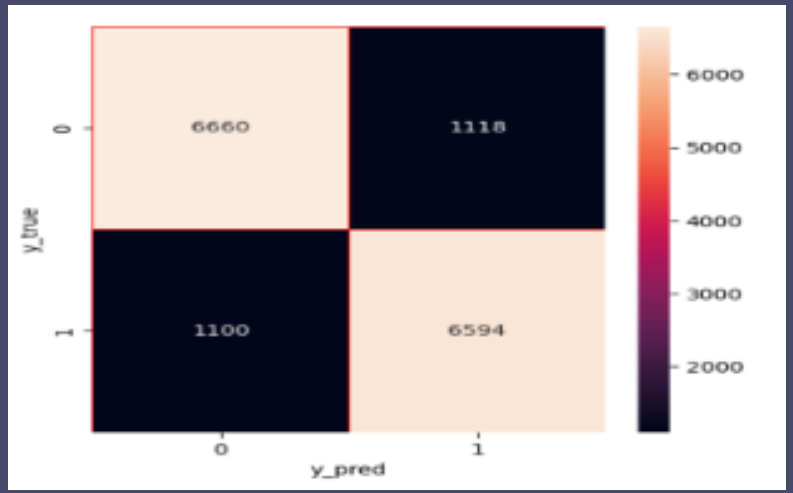
- **Confusion Matrix Of Random Forest**



	precision	recall	f1-score	support
0	0.93	0.97	0.95	7444
1	0.97	0.93	0.95	8028
accuracy			0.95	15472
macro avg	0.95	0.95	0.95	15472
weighted avg	0.95	0.95	0.95	15472

Result & Discussion

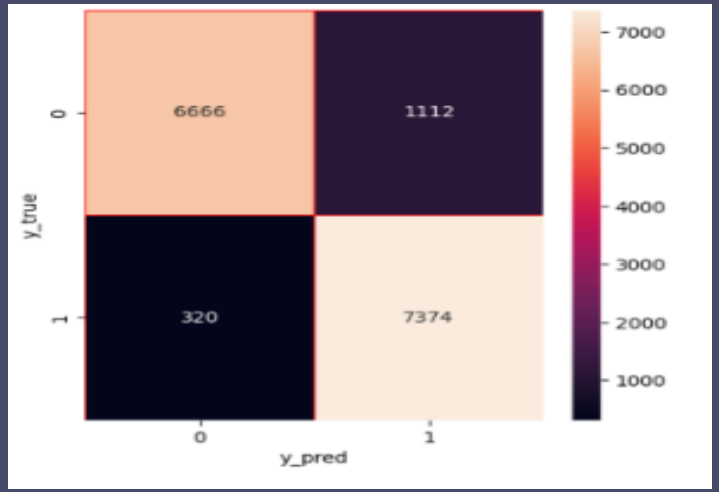
- **Confusion Matrix Of Logistic Regression**



	precision	recall	f1-score	support
0	0.86	0.86	0.86	7760
1	0.86	0.86	0.86	7712
accuracy			0.86	15472
macro avg	0.86	0.86	0.86	15472
weighted avg	0.86	0.86	0.86	15472

Result & Discussion

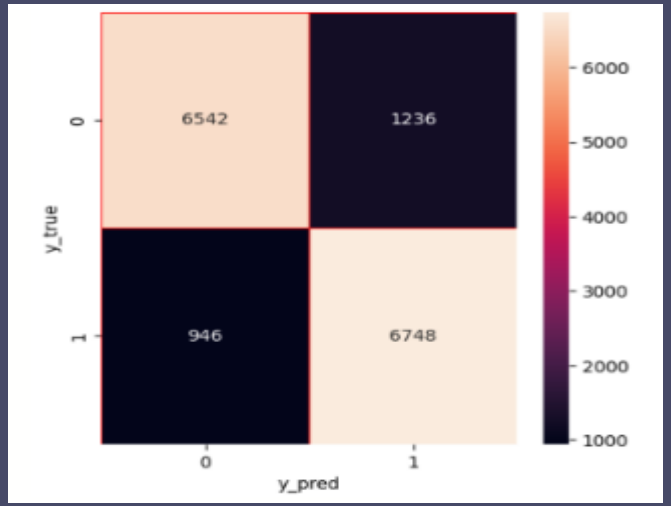
- **Confusion Matrix Of K-Nearest Neighbor**



	precision	recall	f1-score	support
0	0.86	0.95	0.90	6986
1	0.96	0.87	0.91	8486
accuracy			0.91	15472
macro avg	0.91	0.91	0.91	15472
weighted avg	0.91	0.91	0.91	15472

Result & Discussion

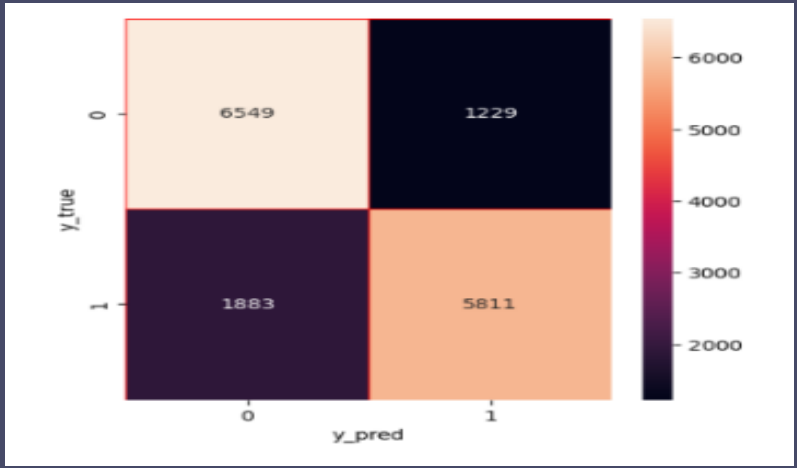
- **Confusion Matrix Of Support Vector Machine**



	precision	recall	f1-score	support
0	0.84	0.87	0.86	7488
1	0.88	0.85	0.86	7984
accuracy			0.86	15472
macro avg	0.86	0.86	0.86	15472
weighted avg	0.86	0.86	0.86	15472

Result & Discussion

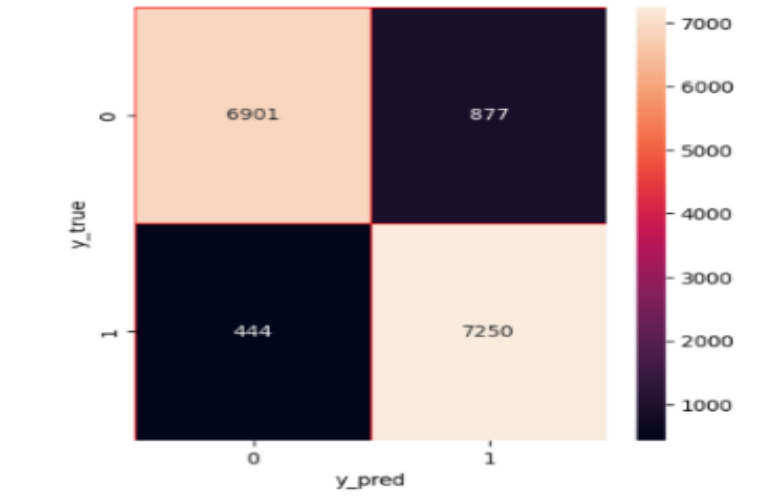
- Confusion Matrix Of Decision Tree



	precision	recall	f1-score	support
0	0.84	0.78	0.81	8432
1	0.76	0.83	0.79	7040
accuracy			0.80	15472
macro avg	0.80	0.80	0.80	15472
weighted avg	0.80	0.80	0.80	15472

Result & Discussion

- **Confusion Matrix Of ANN**



	precision	recall	f1-score	support
0	0.94	0.89	0.91	7778
1	0.89	0.94	0.92	7694
accuracy			0.91	15472
macro avg	0.92	0.91	0.91	15472
weighted avg	0.92	0.91	0.91	15472

Result & Discussion

- Performance Comparison of different Classification algorithms

Parameter	Random Forest	Logistic Regression	Support Vector Machine	K-Nearest neighbors	Decision Tree	ANN	LSTM
Accuracy	**95.2%**	85.6%	85.8%	90.7%	79.8%	91.46%	90.71%
Precision	**95%**	86%	88%	91%	80%	92%	91%
Recall	95%	86%	85%	91%	80%	91%	**99%**
F1-Score	**95%**	86%	86%	91%	80%	91%	95%

Comparison With Previous Work

- Performance Comparison of the target paper dataset and my dataset

Paper Title	Random Forest(RF)	Logistic Regression(LR)	K- Nearest Neighbor(KNN)	Decision Tree(DT)	Artificial neural network(ANN)
Target Paper Dataset	93%	90%	86%	83.12%	88.18%
My Dataset	95.20%	85.66%	91%	79.88%	91.46%

Conclusion

- My research aims to do this by conducting Bank Customer prediction or not as Yes or No.
- Out of the Seven classifiers, i.e. Random Forest (RF), Decision Tree (DT), Logistic Regression (LR),Support Vector Machine(SVM), K-Nearest Neighbors (KNN), Artificial neural network (ANN), and Long short-term memory (LSTM), predictive accuracy of RF is found to be the best.
- We have selected the right model in our paper so that customers get the best prediction and services in the bank.

Future Work

This Work can be extended in the following manner in future

- ❑ **To use more Machine Learning and better hybrid models.**
- ❑ **To use more Deep Learning Models.**
- ❑ **Improve deep learning model accuracy.**
- ❑ **Adding more data.**
- ❑ **Try to use more feature encoding and scaling.**

References

- Syeda Farjana Shetu, Israt Jahan, Mohammad Monirul Islam, Refath Ara Hossain, Nazmun Nessa Moon and Fernaz Narin Nur, “Predicting Satisfaction of Online Banking System in Bangladesh by Machine Learning”, 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST).
- Muhammet Sinan Başarslan, İrem Düzdar Argun, “Prediction of Potential Bank Customers: Application on Data Mining”, Springer Nature Switzerland AG 2020, ICAIAME 2019, LNDECT 43, pp. 96–106, 2020.
- Fereshteh Safarkhani, Sérgio Moro, "Predicting Bank Depositor’s Behavior Using a Decision Tree", Appl. Sci. 2021, 11, 9016. https://doi.org/10.3390/app11199016.
- Sadaf Ilyas, Sultan Zia, Umair Muneer Butt, Sukumar Letchmunan, “Predicting the Future Transaction from Large and Imbalanced Banking Dataset”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020
- Sahar F. Sabbeh, “Machine-Learning Techniques for Customer Retention: A Comparative Study”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018.
- Chaitrali S. Kulkarni, Amruta U. Bhavsar, Savita R. Pingale, Prof. Satish S. Kumbhar, “BANK CHAT BOT – An Intelligent Assistant System Using NLP and Machine Learning”, International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 05 | May -2017.

References

- Regina Esi Turkson, Edward Yeallakuor Baagyere, Gideon Evans Wenya, “A Machine Learning Approach for Predicting Bank Credit Worthiness”, ISBN: 978-1-4673-9187-0 2016 IEEE.
- Mst. Shuly Aktar, “The Impacts of Service Quality on Client Satisfaction: An Empirical Study on Private Commercial Banks in Bangladesh”, Canadian Journal of Business and Information Studies, 3(5), 80-90, 2021.
- Ajmal Amiry, “An Analysis of Customer Satisfaction Level in Banking Sector of Afghanistan”, International Journal of Science and Research (IJSR), Volume 11 Issue 5, May 2022.
- Chengming Chang, “Research on Domestic Bank Customer Satisfaction Based on Logistic Regression Analysis”, International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2015).
- Dataset name “Bank Customer Data”. Available : https://www.kaggle.com/datasets/kidoen/bank-customers-data.

Thank You