TD Predictor

TERM DEPOSIT ANALYSIS AND PREDICTION

EDA, DATA PREPROCESSING AND COMPARING DIFFERENT MACHINE LEARNING MODELS FOR PREDICTING SUBSCRIPTION TO BANK TERM DEPOSIT.

TD Predictor

TEAM MEMBERS:

- UTKARSH MISHRA (11600119031)
- PRATYUSH MAJUMDER (11600119046)
- ROSHNI DEY (11600119033)
- TANISHA GHOSH (11600119050)
- TIYASHA PARUI (11600119028)

GROUP ~ 6
MENTOR ~ Ms. RACHITA GHOSHHAJRA
COORDINATOR ~ Mr. SUMIT MAJUMDAR

TD Predictor

ACKNOWLEDGEMENT:

- We would take the opportunity to express our deep sense of gratitude to **Mr. Avijit Bose,** Head of the Department, Computer Science and Engineering, MCKV Institute of Engineering for supporting us to make our project worth it.
- We would want to convey our thanks to **Ms. Rachita Ghoshhajra**, our mentor, for guiding us through this process. We've learned a lot of new things thanks to her guidance, and we're looking forward to learning even more in the future.
- The project coordinator**, Mr. Sumit Majumdar,** is also to be thanked for providing and delineating the administrative procedures pertaining to project processes.

TD Predictor

BACKGROUND:

- **Machine learning** is a broad term encompassing a number of methods that allow the investigator to learn from the data. These methods may permit large real-world databases to be more rapidly translated to applications to inform patient-provider decision making.
- A wide variety of approaches, algorithms, statistical software, and validation strategies were employed in the application of machine learning methods to inform patient-provider decision making.
- There is a need to ensure that multiple machine learning approaches are used, the model selection strategy is clearly defined, and both internal and external validation are necessary to be sure that decisions for patient care are being made with the highest quality evidence.
- Our model will include Data preprocessing and Comparing different machine learning models for predicting subscription to **bank term deposit**.

TD Predictor

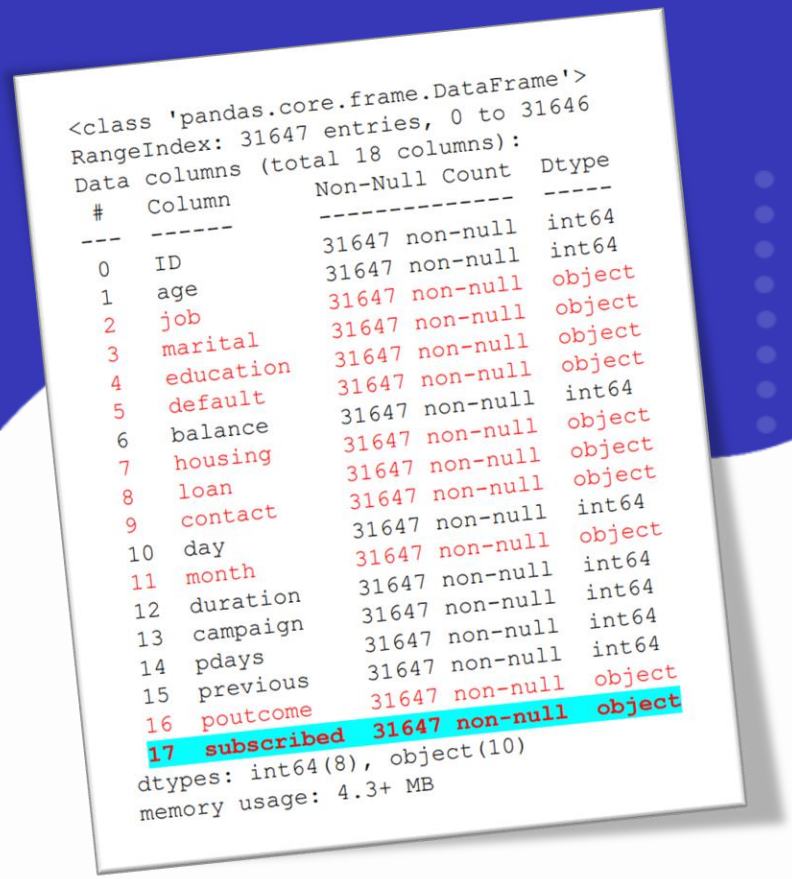
INTRODUCTION:

- A **term deposit** is a cash investment held at a financial institution. Your money is invested for an agreed rate of interest over a fixed amount of time, or term.
- The bank has various outreach plans to sell term deposits to their customers such as email marketing, advertisements, telephonic marketing and digital marketing.
- Telephonic marketing campaigns still remain one of the most effective way to reach out to people. However, they require huge investment as large call centers are hired to actually execute these campaigns.
- Hence, it is crucial to identify the customers most likely to convert beforehand so that they can be specifically targeted via call.

TD Predictor

INTRODUCTION TO DATASET:

PROPERTY	DATASET 1	DATASET 2
ENTRIES	31647	45211
VARIABLES	18	18
TARGET VARIABLE	subscribed	subscribed
INDEPENDENT VARIABLES	17	17
DEPENDENT VARIABLE	01	01
CONTINUOUS VARIABLES	10	10
CATEGORICAL VARIABLES	08	08
SHAPE	(31647,18)	(45211,18)



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31647 entries, 0 to 31646
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   ID           31647 non-null  int64
 1   age          31647 non-null  int64
 2   job          31647 non-null  object
 3   marital      31647 non-null  object
 4   education    31647 non-null  object
 5   default      31647 non-null  object
 6   balance      31647 non-null  int64
 7   housing      31647 non-null  object
 8   loan         31647 non-null  object
 9   contact      31647 non-null  object
 10  day          31647 non-null  int64
 11  month        31647 non-null  object
 12  duration     31647 non-null  int64
 13  campaign     31647 non-null  int64
 14  pdays        31647 non-null  int64
 15  previous     31647 non-null  int64
 16  poutcome     31647 non-null  object
 17  subscribed   31647 non-null  object
dtypes: int64(8), object(10)
memory usage: 4.3+ MB
```

TD Predictor

WORKFLOW MODEL:

Data Collection

Data collection is the process of gathering and selecting relevant data to be used as input for training and testing machine learning models.

Importing required libraries and dataset

Data collection is the process of gathering and selecting relevant data to be used as input for training and testing machine learning models.

Exploratory Data Analysis

EDA (Exploratory Data Analysis) is the process of analyzing and visualizing data to gain insights and discover patterns, relationships, and anomalies.

Data Pre Processing

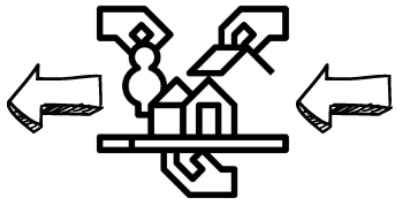
Data pre-processing involves cleaning, transforming, and formatting data to make it suitable for machine learning algorithms.

Dimentionality Reduction

Data dimensionality reduction is the process of reducing the number of input features while retaining the most relevant information to improve the efficiency and accuracy of machine learning models.

Data Splitting

Test-train split is a technique used in machine learning to divide a dataset into two separate sets, one for training a model and the other for evaluating its performance.



Model Building

Model building is the process of selecting a suitable algorithm, training it on a dataset, and fine-tuning it to achieve optimal performance for a specific task.

Stacking and Export

Stacking is a machine learning technique that combines multiple models to improve the accuracy of predictions, and once the model is trained, it can be exported and deployed to make predictions on new data.

Frontend Integration

Frontend integration involves incorporating machine learning models into a user-facing application or website to provide personalized experiences or insights for users.

Model Ready

A model is ready when it has been trained and fine-tuned to achieve optimal performance for a specific task and is ready to be deployed for making predictions on new data.

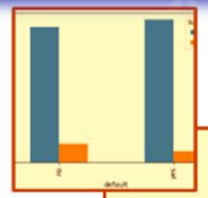
TD Predictor

EXPLORATORY DATA ANALYSIS:

- EDA stands for **Exploratory Data Analysis** and is a crucial step in data analysis.
- It involves summarizing and visualizing data to identify patterns, relationships, and outliers.
- EDA is an iterative process that helps analysts to better understand the data they are working with.
- It can identify potential issues with the data, such as missing values or data entry errors.
- It is an essential step in the data analysis process and can help analysts uncover valuable insights.

Univariate Analysis

- Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it.



Bivariate Analysis

- Bivariate analysis is where you are comparing two variables to study their relationships. These variables could be dependent or independent to each other. In Bivariate analysis is that there is always a Y-value for each X-value.

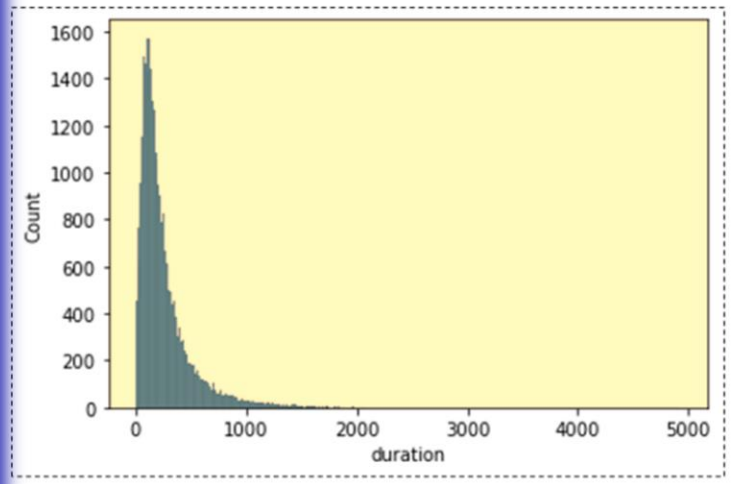
Multivariate Analysis

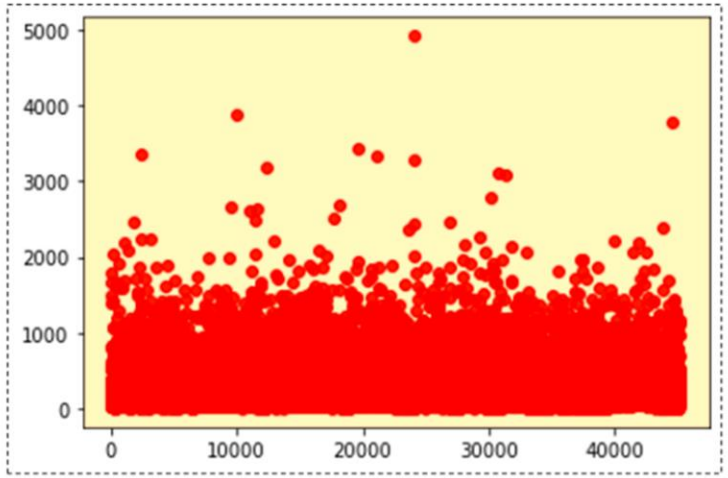
- Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables. For three variables, However, since we cannot visualize anything above the third dimension, we often rely on other softwares and techniques for us to be able to grasp the relationship in the data.

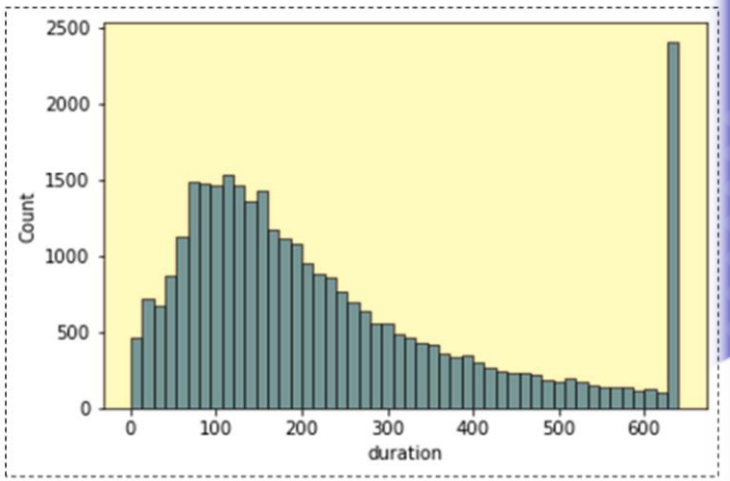
TD Predictor

DURATION – EDA AND OUTLIER TREATMENT:

Scatter plot visualization







Outlier treatment

Distribution of duration values showing skewness towards right and long x axis showing the possible presence of outliers

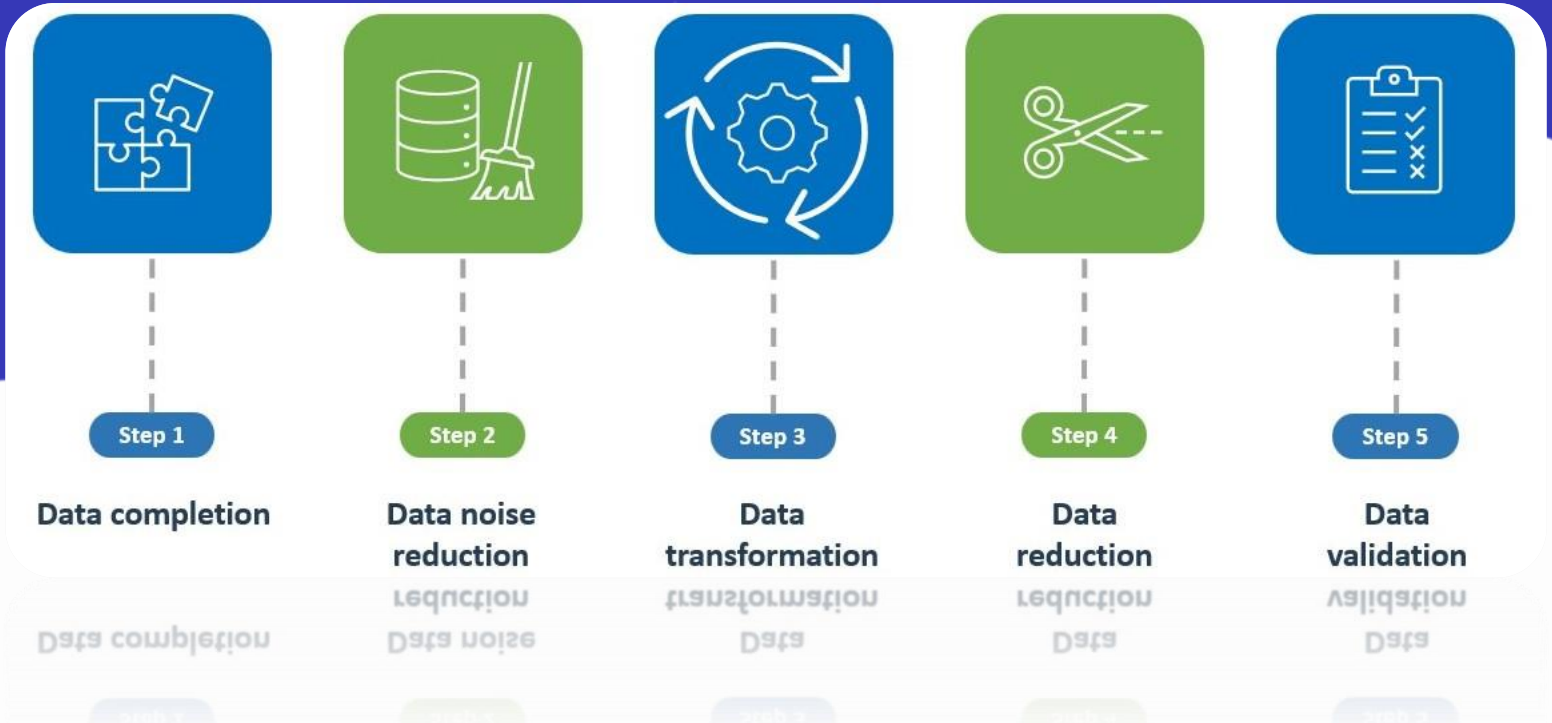
A scatter plot confirming the presence of outliers in the dataset that require further treatment to decrease dataset complexity

Resultant distribution being better than before with excessively large and small values being taken care of using the method of inter quantile range.

TD Predictor

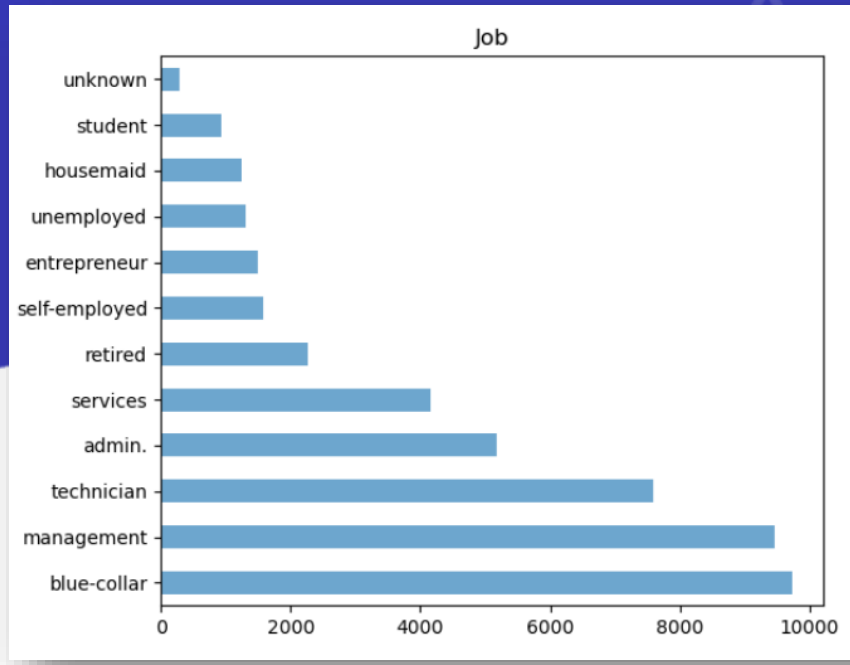
DATA PRE-POCESSING:

- **Data preprocessing** is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.
- Why do we need Data Preprocessing?
  - Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model
  - Steps involved in data pre-processing are as follows:



CATEGORICAL VARIABLE CONVERSION:

- **Extensive data processing** can be a difficult and time-consuming operation that calls for specialized knowledge and software solutions. However, the understanding that can be attained from delving into vast amounts of data can result in appreciable improvements in organizational performance, effectiveness, and decision-making
- It can also be observed that certain jobs like **entrepreneur, blue-collar, services and housemaid** share same percentage of subscribed individuals and it is way more less when compared with jobs of other categories. Hence, they can be grouped together while converting them into category based numeric form for them to be categorized as a single entity of job



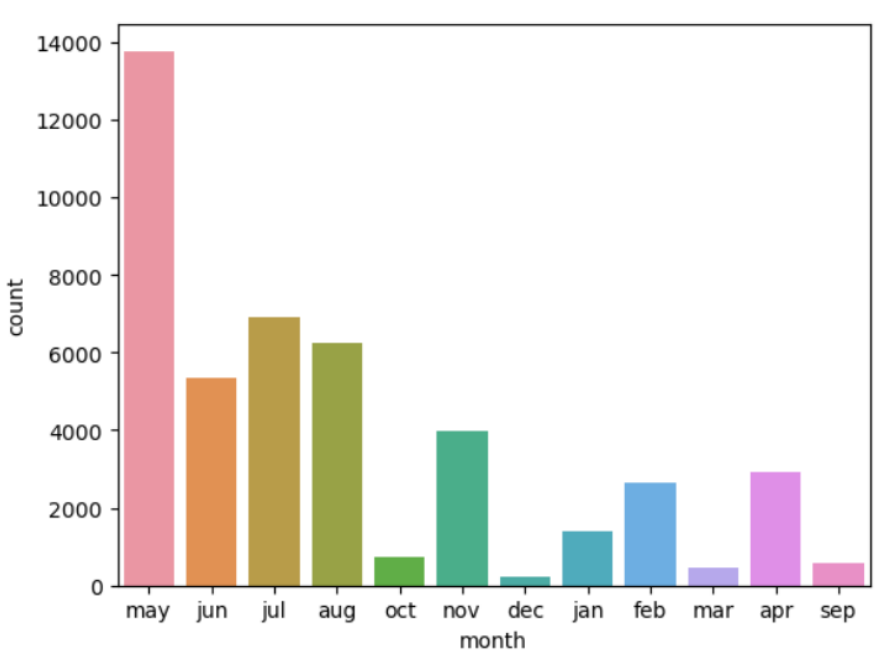
job	subscribed
student	0.286780
retired	0.227915
unemployed	0.155027
management	0.137556
admin.	0.122027
self-employed	0.118429
unknown	0.118056
technician	0.110570
services	0.088830
housemaid	0.087903
entrepreneur	0.082717
blue-collar	0.072750

JOB	VALUE
Unknown	5
Student	9
Housemaid	1
Unemployed	7
Entrepreneur	1
Self Employed	4
Retired	8
Services	1
Admin	3
Technician	2
Management	6
Blue Collar	1

TD Predictor

CATEGORICAL VARIABLE CONVERSION:

- Similar to the type of conversion performed for job, it can be said that certain months, including **May, June, July, January, and November**, share a similar percentage of individuals who have subscribed, and this percentage is significantly lower than that of other job categories.
- Therefore, these months can be grouped together and converted into a category-based numeric form to be categorized as a single job entity. The correlation before and after grouping and numerical transformation was 2% and 13.5%, respectively.



MONTH	CONVERSION
January	01
February	02
March	03
April	04
May	05
June	06
July	07
August	08
September	09
October	10
November	11
December	12

Correlation = 01.87%

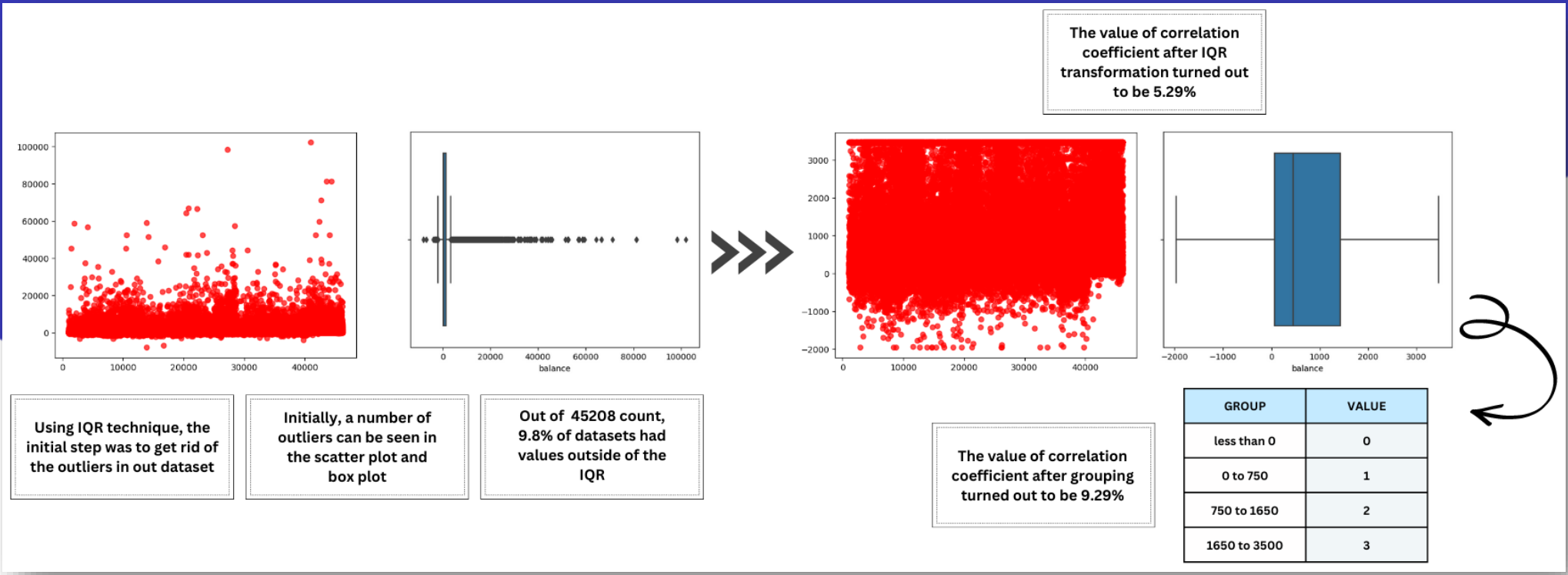
MONTH	CONVERSION
January	01
February	02
March	03
April	04
May	01
June	01
July	01
August	08
September	09
October	10
November	01
December	12

Correlation = 13.51%

TD Predictor

CONTINUOUS VARIABLE CONVERSION:

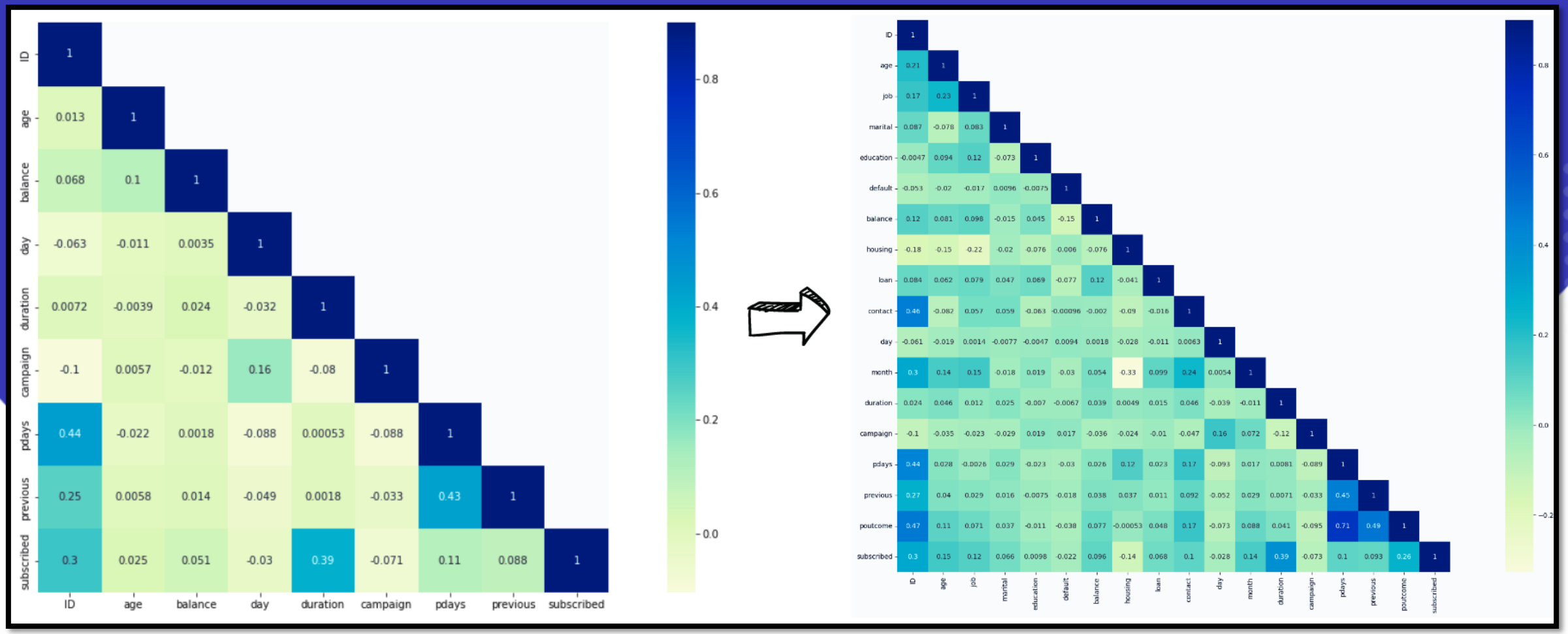
- The variable **"balance"** is a numerical variable that can be used for modeling purposes, but its complexity can be reduced by removing outliers or grouping the data into bands. At first, "balance" had a correlation value of 0.05283 (or 5%) with the target variable.
- However, after applying outlier removal and grouping the data into groups of four, the correlation value increased considerably to 0.09617 (or approximately 10%).



GROUP	VALUE
less than 0	0
0 to 750	1
750 to 1650	2
1650 to 3500	3

TD Predictor

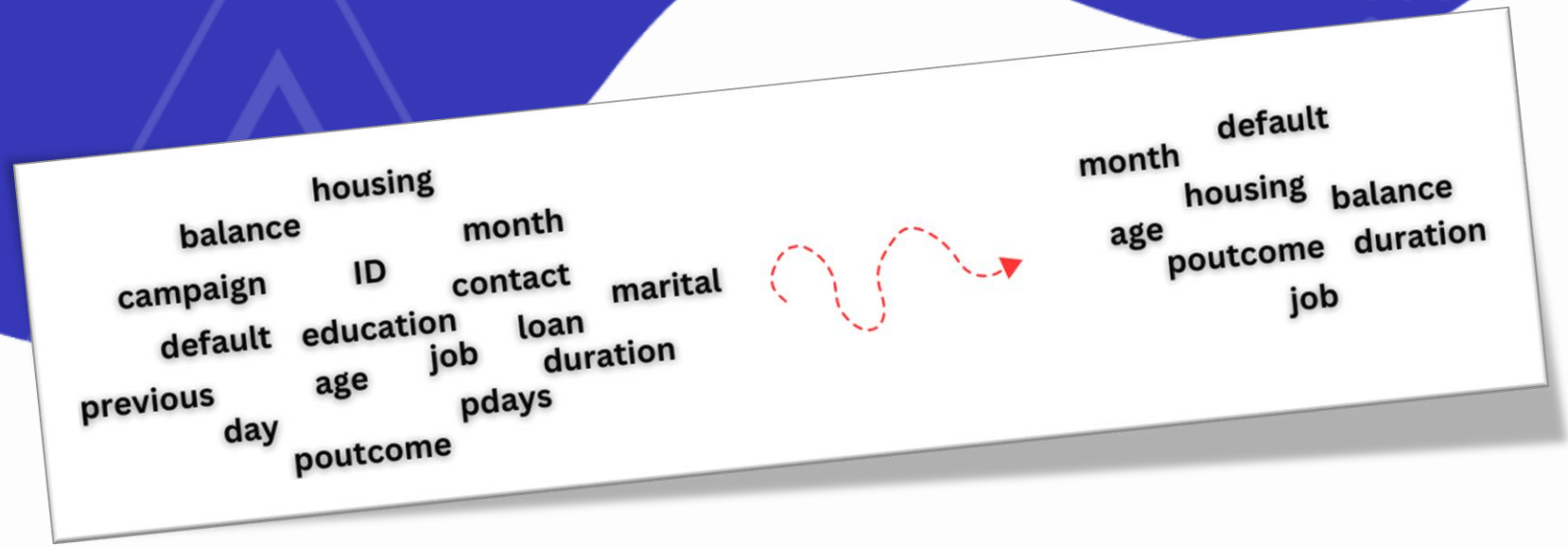
UPGRADED CORRELATION HEATMAP:



TD Predictor

FEATURE SELECTION:

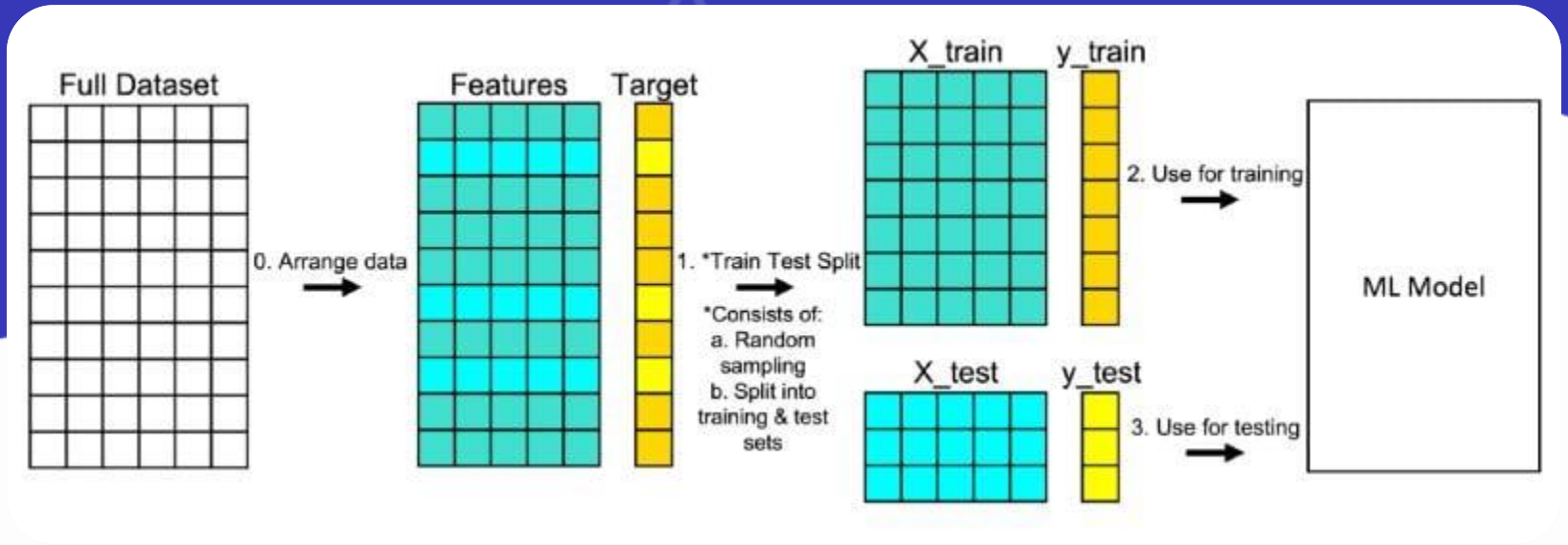
- **Feature selection** is the process of selecting a subset of relevant features from a larger set of features in a dataset.
- The goal of feature selection is to improve the performance of a model by reducing the dimensionality of the input space and removing irrelevant or redundant features.
- Feature selection can improve model accuracy, reduce overfitting, and speed up training times. It is a critical step in machine learning and can help improve model performance and efficiency.
- In our model, we utilized backward elimination as a method to remove variables that displayed multicollinearity, had a low correlation coefficient, and were deemed unimportant for the process of constructing the model.



TD Predictor

TEST TRAIN SPLIT:

- The **train-test split** procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.
- If your training accuracy is high, but your testing accuracy is low, you can't really advertise your model as a good model.
- **Cross-validation** is a resampling method that tests and trains a model on different iterations using different chunks of the data.

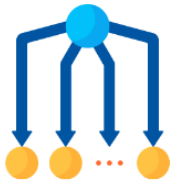


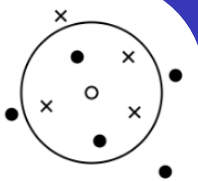
TD Predictor

IMPLEMENTATION:

ALGORITHMS USED:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Naive Bayes
5. K Nearest Neighbours





ML ALGORITHM	LOGISTIC RERESSION	DECISION TREE	RANDOM FOREST	NAIVE BAYES	KNN
IMPORTING REQUIRED LIBRARY	from sklearn.linear_model import LogisticRegression	from sklearn.tree import DecisionTreeClassifier	from sklearn.ensemble import RandomForestClassifier	from sklearn.naive_bayes import GaussianNB	from sklearn.neighbors import KNeighborsClassifier
DEFINING THE MODEL	lreg = LogisticRegression()	dtree = DecisionTreeClassifier(criterion='gini', max_depth=6, random_state=0)	rfc = RandomForestClassifier(n_estimators=6, random_state=0, max_features=6, max_depth=6)	classifier = GaussianNB()	knn = KNeighborsClassifier(n_neighbors=7)
FITTING THE MODEL ON TRAINING DATASET	lreg.fit(x_train,y_train)	dtree.fit(x_train, y_train)	rfc.fit(x_train, y_train)	classifier.fit(x_train, y_train)	knn.fit(x_train, y_train)
MAKING PREDICTION	lr_pred = lreg.predict(x_test)	dt_pred = dtree.predict(x_test)	rf_pred = rfc.predict(x_test)	nb_pred = classifier.predict(x_test)	knn_pred=knn.predict(x_test)
PERFORMANCE CHECK	accuracy_score(y_test, lr_pred)	accuracy_score(y_test, dt_pred)	accuracy_score(y_test, rf_pred)	accuracy_score(y_test, nb_pred)	accuracy_score(y_test, knn_pred)

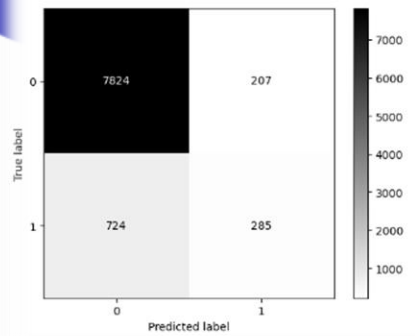
TD Predictor

FINAL RESULTs:

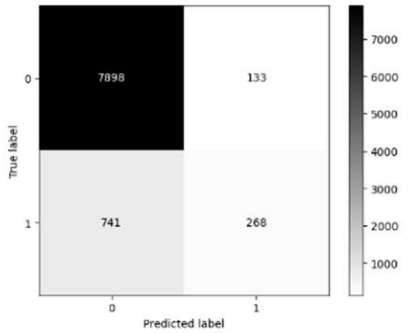
EXPERIMENTAL RESULTS:

1. Accuracy
2. Precision
3. Recall
4. F1 Score
5. Micro Average
6. Weighted Average
7. Cross Validation

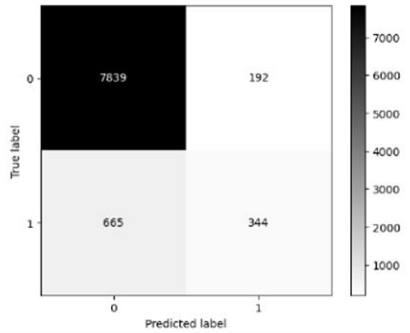
PARAMETERS	LOGISTIC RERESSION	DECISION TREE	RANDOM FOREST	KNN	NAIVE BAYES
ACCURACY	90%	90%	91%	89%	89%
PRECISION	0 - 92% 1 - 58%	0 - 91% 1 - 67%	0 - 92% 1 - 64%	0 - 91% 1 - 51%	0 - 93% 1 - 49%
RECALL	0 - 97% 1 - 28%	0 - 98% 1 - 27%	0 - 98% 1 - 34%	0 - 98% 1 - 20%	0 - 95% 1 - 40%
F1 SCORE	0 - 94% 1 - 38%	0 - 95% 1 - 38%	0 - 95% 1 - 65%	0 - 94% 1 - 29%	0 - 94% 1 - 44%
MICRO AVERAGE	Precision: 75% Recall: 63% F1-Score: 66%	Precision: 79% Recall: 62% F1-Score: 66%	Precision: 78% Recall: 66% F1-Score: 70%	Precision: 71% Recall: 59% F1-Score: 61%	Precision: 71% Recall: 68% F1-Score: 69%
WEIGHTED AVERAGE	Precision: 88% Recall: 90% F1-Score: 88%	Precision: 98% Recall: 90% F1-Score: 88%	Precision: 89% Recall: 91% F1-Score: 89%	Precision: 86% Recall: 89% F1-Score: 67%	Precision: 88% Recall: 89% F1-Score: 88%
CROSS VALIDATION	89% 93% 91% 87% 88%	88% 89% 90% 88% 86%	89% 88% 89% 85% 87%	88% 88% 87% 88% 86%	92% 89% 91% 88% 90%



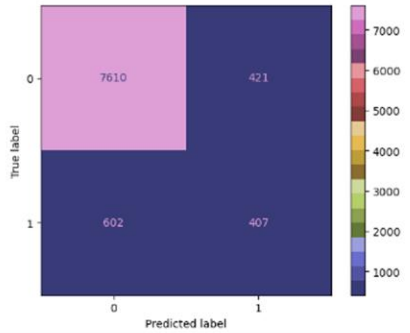
LR



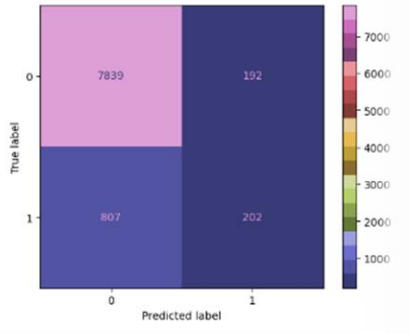
DT



RF



NB

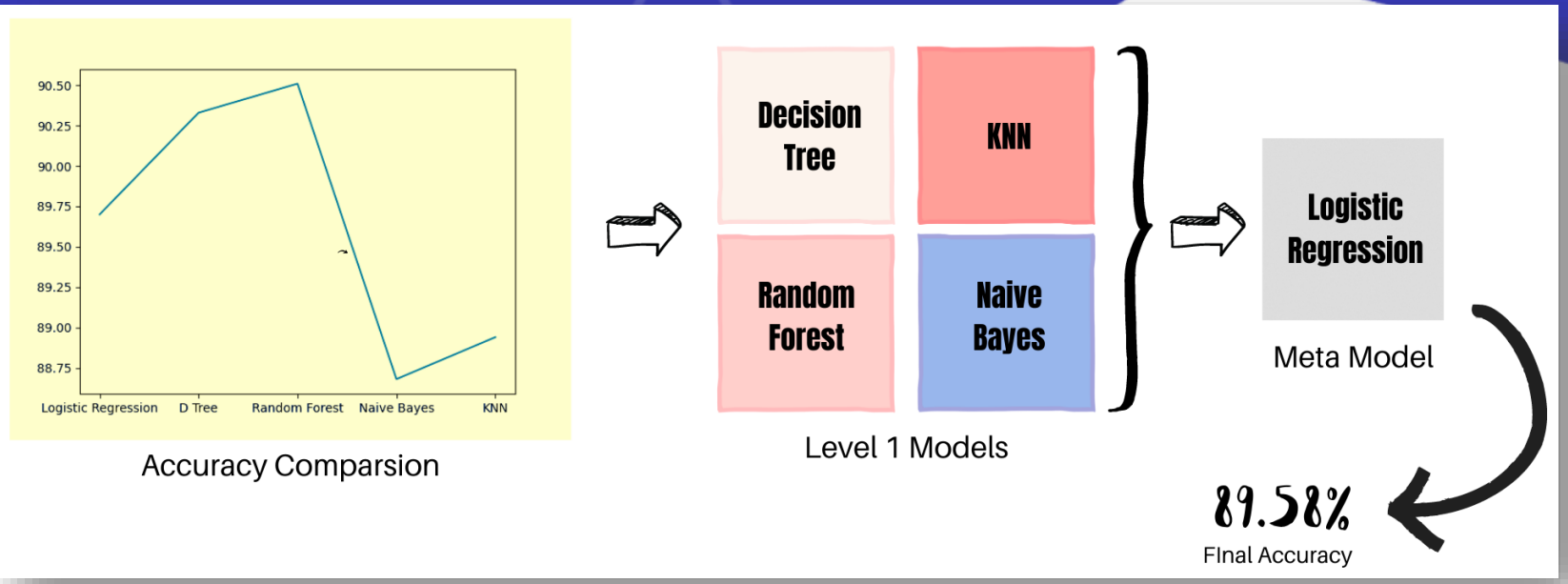


KNN

TD Predictor

STACKING:

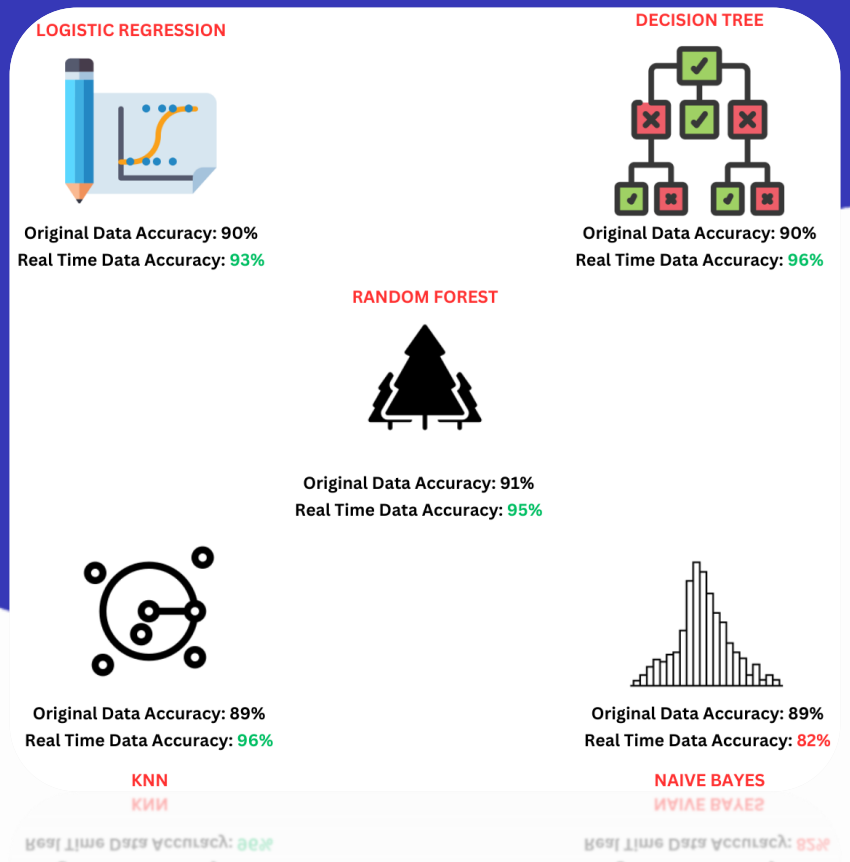
- **Stacking**, also known as stacked generalization, is a meta-learning ensemble technique that involves combining multiple models to improve prediction accuracy.
- In stacking, the outputs of several base models are used as input features to a higher-level "meta-model", which makes the final predictions. The base models can be trained on the same or different datasets and can use different learning algorithms.
- "By achieving approximately 90% accuracy across different models, we can infer that we have significantly simplified our dataset and successfully conducted thorough data cleaning processes."



TD Predictor

WORKING ON REAL TIME DATASET:

- We have gathered over **450 real-time data entries** from various individuals. Subsequently, we applied **exploratory data analysis** (EDA) and **data pre-processing** techniques to transform the dataset. Finally, we utilized the processed dataset to evaluate our model, leading to the following observations:



TD Predictor

CONCLUSION:

- In this entire money exchange process and banks undergoing digitalization, our work will assist banks in determining the clients who are most likely to opt for a term deposit plan, allowing banking institutions to save a significant amount of spending.
- For our initial dataset, accuracy of random forest turned out to be the best (91%), followed by Naïve Bayes with the highest precision value (94%) whereas identifying the actual positives that our model correctly predicted was shown by random forest with the highest recall value (98%).
- Real time dataset showed better results and accuracy when used for testing.

TD Predictor

FUTURE SCOPE:

- Model Improvement
- Scalability and Development with larger datasets
- Continuous learning and adaptation
- Integration with other technologies

TD Predictor

REFERNECES:

[1] S. Manlangit, S. Azam, B. Shanmugam and A. Karim," Novel Machine Learning Approach for Analysing Anonymous Credit Card Fraud Patterns", International Journal of Electronic Commerce Studies 10.2 (2019): 175-202

M. Sergio, L. Raul and C. Paulo," Using data mining for bank direct marketing: an application of the CRISP-DM methodology", RepositoriUM, 2011

C. S. T. Koumetio, W. Cherif and S. Hassan," Optimizing the prediction of telemarketing target calls by a classification technique," 2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM), 2018

J. Asare-Frempong and M. Jaya Balan," Predicting customer response to bank direct telemarketing campaign," 2017 International Conference on Engineering Technology and Technopreneur ship (ICE2T), 2017

TD Predictor

REFERNECES:

Rony MA, Hassan MM, Ahmed E, Karim A, Azam S, Reza DA. Identifying Long-Term Deposit Customers: A Machine Learning Approach. In2021 2nd International Informatics and Software Engineering Conference (IISEC) 2021 Dec 16 (pp. 1-6). IEEE.

Bisong E. Matplotlib and seaborn. InBuilding machine learning and deep learning models on google cloud platform 2019 (pp. 151-165). Apress, Berkeley, CA.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.

Song YY, Ying LU. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 2015 Apr 4;27(2):130.

TD Predictor

REFERNECES:

A. Gupta, A. Raghav and S. Srivastava," Comparative Study of Machine Learning Algorithms for Portuguese Bank Data," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021.

G. Sell and D. Garcia-Romero," Diarization resegmentation in the factor analysis subspace," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4794-4798

A. Sarmento, K. Yeo, S. Azam, A. Karim, A. Al Mamun and B. Shanmugam," Applying Big Data Analytics in DDos Forensics: Challenges and Opportunities", Cybersecurity, Privacy and Freedom Protection in the Connected World, pp. 235-252, 2021.

TD Predictor

THANK YOU