**Predicting Customer Response To Bank Marketing Campaigns Using Machine Learning**



PROJECT REPORT SUBMITTED TO

Symbiosis Institute of Geoinformatics

FOR PARTIAL FULFILLMENT OF THE M.Sc. DEGREE

By

***SOURAV BALASAHEB KHOT***

**PRN 22070243027**

**M.Sc. (Data Science and Spatial Analytics)**

**BATCH 2022-24**

Symbiosis Institute of Geoinformatics

Symbiosis International (Deemed University)

5th Floor, Atur Centre

Gokhale Cross Road

Model Colony

Pune – 411016

Maharashtra

India

February 2023

**Acknowledgment**

I would like to convey my heartfelt gratitude to **Dr. Vidya Patkar** and **Mr. Sahil Shah** for their tremendous support and assistance in completing my project. I would also like to thank them, for providing me with this beautiful opportunity to work on a project with the topic “**Predicting Customer Response To Bank Marketing Campaigns Using Machine Learning**” The completion of the project would not have been possible without their help and insights.

I am grateful to all those with whom I have had the pleasure to work during this and other related projects. Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. They are the ultimate role models.

**Sourav Khot**

**Contents :**

|  |  |  |
| --- | --- | --- |
| **Sr No.** | **Topics** | **Page No.** |
| **1.** | Abstract | 4 |
| **2.** | Introduction | 4 |
| **3.** | Dataset Details | 6 |
| **4.** | Libraries | 9 |
| **5.** | Data cleaning and preprocessing | 10 |
| 6. | Model Design | 16 |
| **7.** | Visualization | 21 |
| **8.** | Conclusion | 22 |
| **9.** | References | 23 |

**Abstract:**

Achieving growth is a key objective for any banking organization, and several strategies can be employed to achieve this goal. For planning the strategies, we have to analyze or we need to perform some operations on data. In this report, we are going to see some trends and patterns of data that help to make decisions, and for that, we used different machine learning models. That gives us insights from the data which is one of the main motives of this project. This bank Marketing dataset contains information on direct campaigns (phone calls). The objective of the campaign is to promote bank term deposits and the data set provides information on whether or not the clients subscribed to the product. In these projects, we performed exploratory data analysis (EDA) to understand the data, visualization, and machine learning models such as logistic regression, and random forest, to build predictive and classification models to evaluate the performance.

**Introduction:**

In these reports, we have collected data from the Kaggle site where data contains different features like age, job, marital status, education, contact, and duration that helps us to decide whether the customer will subscribe to termed deposit or not. The termed deposit is a process where customer stores their payment or saves money for a fixed time and after that money, the bank will give some additional amount. In banking terms, we say it also as a fixed deposit policy. This research may aim to create a machine-learning model that can forecast whether or not a customer would open a term deposit, a kind of long-term savings account. The project can entail investigating the data to identify trends and connections among the various variables, preparing the data to make it appropriate for analysis, and developing and testing a prediction model using various machine-learning approaches. In these projects, we have used different libraries like psycopg2 to connect the python script with the database, and for visualization means dynamic visualization we used libraries like Plotly. Also, we used a library like pandas\_profiling for preprocessing.

These data are related to the Portuguese bank system. Where different factors like the customer’s age, Job details, customer’s marital status, if the customer has taken a loan or not, and contact means how many days ago contacted the customer all these factors matter to predict whether the customer will subscribe to a term deposit or not. For that, we used models like logistic regression, random forest, KNN (k nearest neighbor) classifier, and support vector machine(SVM). After that, we have done some visualization with the help of Matplolib, Seaborn, and Plotly library. from visualization some things are cleared and easily visualize from that we can find some insights that will be done in these project.



calls) of a Portuguese banking institution. The objective of the campaign was to promote bank term deposits, and the dataset provides information on whether or not the clients subscribed to the product. The dataset "Bank Marketing" contains information on direct marketing campaigns (phone calls)a Portuguese bank term deposits, and the dataset provides information on whether o**r not the clients subscribed to the**

**Dataset Details:**

**Data source:**

This data is collected from the Kaggle datasets site.

**Dataset Name:**

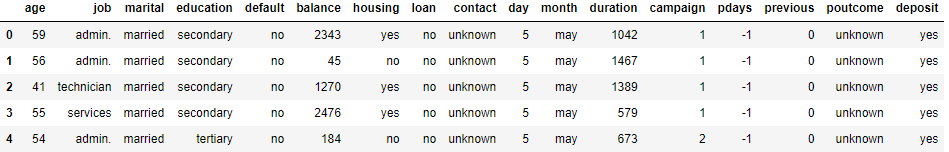
Bank Marketing

**Source**: [https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing](%20https:/www.kaggle.com/datasets/henriqueyamahata/bank-marketing)

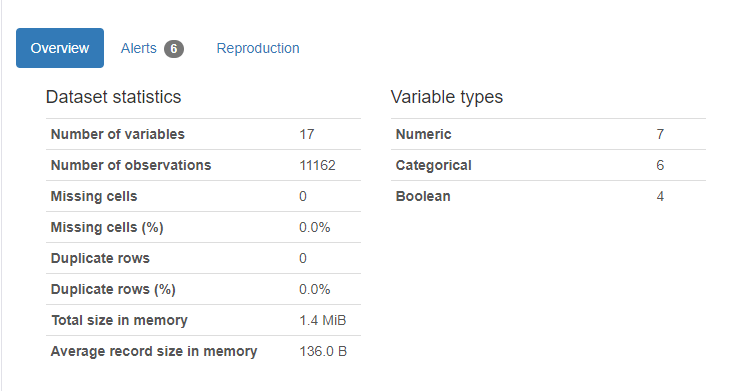
**Data Shape:** 11162 X 17

This data is Dynamic data because if a new customer created a bank account and if they want to deposit their money in the bank that also will be added to these reports

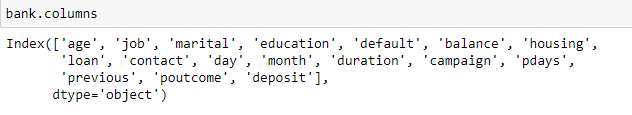
**Data Sample:**

.

Here is an overview of the data drawn with the help of the pandas\_ profiling library:



**Dataset columns:**

****

here are these different data columns are present every column has its own identity that as follows:

1. **Age**: It shows the age of the customer This is a numeric column.
2. **job**: It shows the type of job or description of the job of the customer it is a categorical column.
3. **marital**: It shows the marital status of the customer; it is a categorical column.
4. **Education**: It shows the education level of the customer; it is a categorical column.
5. **Default**: It shows whether the customer has credit in default or not, it is a categorical column.
6. **Balance:** It shows the average yearly balance in Euros of the customer, it is a numeric column.
7. **Housing**: It shows whether the client has taken a housing loan or not, it is a categorical column.
8. **Loan:** It shows whether the client has another personal loan or not, it is a categorical column.
9. **contact:** It shows the contact communication type, it is a categorical column.
10. **Day**: It shows the last contact day, it is a numeric column.
11. **month:** It shows the last contact month, it is a categorical column.
12. **Duration:** It shows the last contact duration and how many seconds talk on a call, it is a numeric column.
13. **Campaign**: It shows the number of contacts performed for this client, it is a numeric column.
14. **Pdays**: It shows the number of days the last contact was done (-1 means the client was not previously contacted), it is a numeric column.
15. **Previous**: It shows the number of contacts done with this customer before this campaign, it is a numeric column.
16. **poutcome:** It shows some previous year’s outcomes of marketing campaigns, it is a categorical column.
17. **Deposite**: It shows whether the client subscribed to a term deposit or not, it is binary data where values like yes or no are present.

**Libraries:**

Here are some libraries used in these project

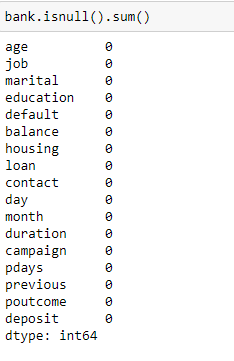
* Pandas:- This library is useful for reading data in .csv format.
* Numpy:- This library is used for doing some numeric operations in data.
* Matplotlib:- This library is used for visualizing the data their sub-package is used that

named a pyplot from which we can draw different graphs.

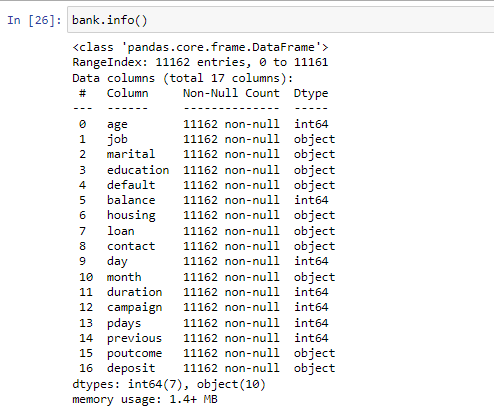
* Seaborn:- This library is also used for visualization.
* Psycopg2:- This library is used to connect databases with python or a server.
* Warnings:- To remove the warning message
* Plotly:- It most useful library for visualization from which we can create a dynamic visualization plot.
* Pandas\_profiling: This library is useful for the preprocessing purpose from that we can get information about every attribute visually.
* IPython.display: This library is useful for taking images from websites and attaches with the help URL in a notebook.
* Sklearn.model\_selection: from sci-kit learn we have to select different models like train\_test\_split and other metrics related and accuracy scores.

**Data Cleaning and Preprocessing:**

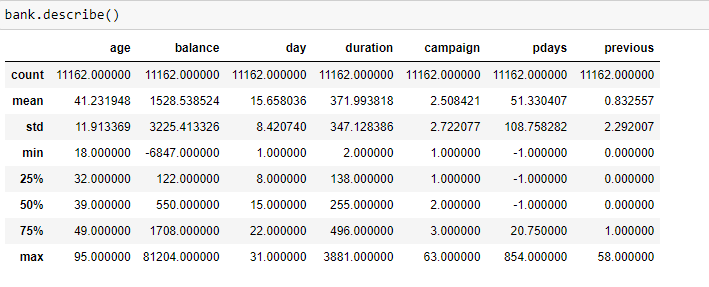
For getting more information on data like data size and data shape we used the shape and size function. This dataset is collected from the Kaggle datasets website there will be no missing values found so there will be no need to handle missing values.



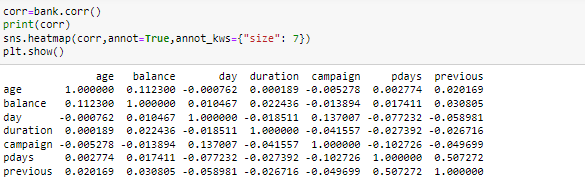
For checking the datatype of the column we used the info() function

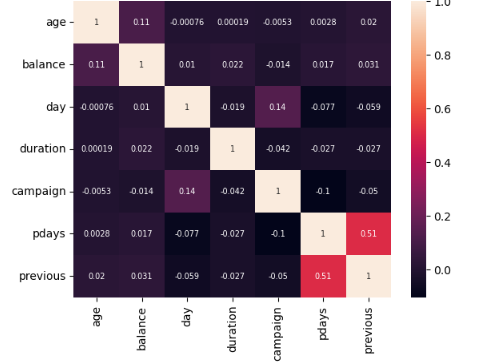


For checking values in the data or statistical values in data we used the following queries.



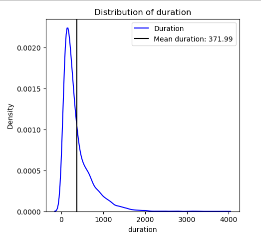
For checking the relationship between two variables or how variables are correlated with each other we used the correlation function and also for that we plot the correlation visually with the help of a seaborn heatmap.





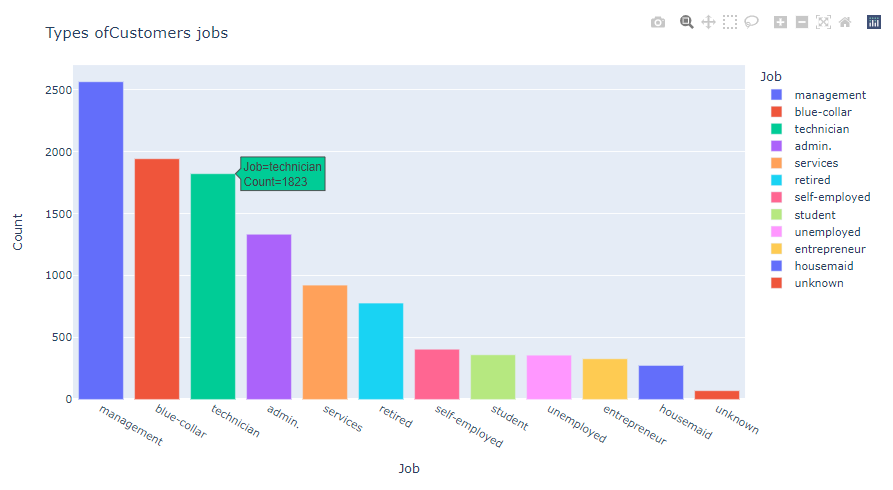
Also drew a seaborn pair plot to visualize the data correlation using graphs from these we can view data on how much scattered and how the correlation goes whether it is positive or negative.

From correlation, we can see that the duration column is more differ in range for that also drawn a normal distribution type curve



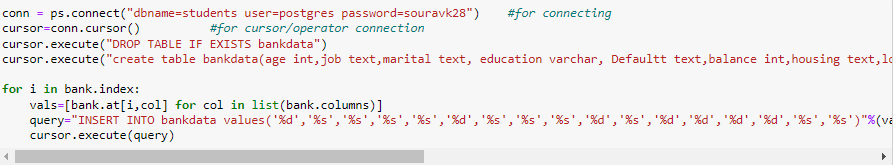
To check how many types of customer jobs are done with the help of the Plotly graph as follows:





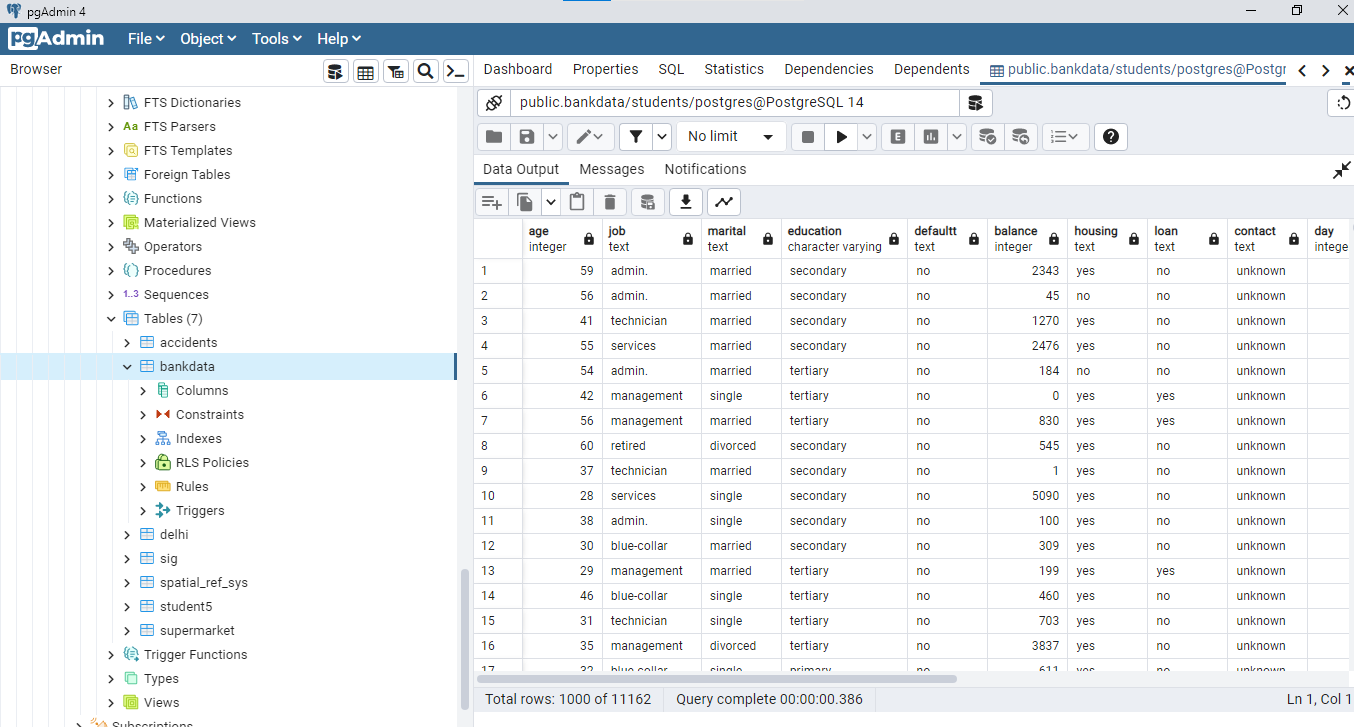
Also for checking the outliers in the column of marital status versus age, we use a Plotly box-plot that helps analyze the result there is some outlier personality visualize like the marital status of age 86 person is coming single that will be easily visualized with this graph that helps to analyze and built the model.

For running the model, we have connected the data with the bank data database with the help of the psycopg2 model for that code is below:



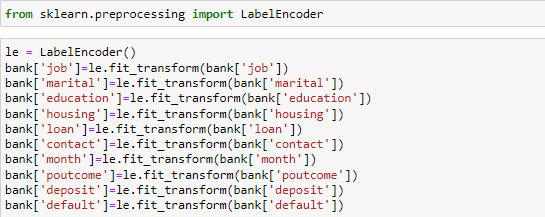
Here it is important to store that amount of data in the database because it is dynamic data if a new customer comes their data should be stored in the database it is practically not possible to store data in the file format or any hard disk format because we have to give the access of that data to other users for simultaneously working on these data for that purpose we used to store the data in cloud storage or database storage in these case we stored the data in ‘student’ database with the table name ‘bankdata’.from these many persons access the data with DCL commands and operate on it for that owner wants to permit for that queries like grant and revoke are present also we can do different operations on it like Insert, update, delete, truncate and many more this is the DML operations.

Here is the PgAdmin4 view where data is stored in the ‘student’ database



As these data have stored in the database and also we can perform various operations on it.

After that, for the model building, we need to convert the categorical values into numerical values for that we need to use encoding there are different types of encoding like hot encoding, and label encoding here we are using label encoding for different features as follows.



After that, we will perform the different models.

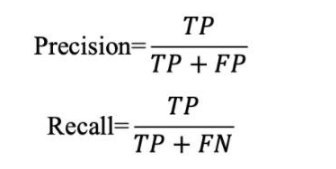
**Model Design:**

Model designing is one of the important processes of any machine learning project, From the data, we need to decide the input variables and output variables here in these data our main problem statement is finding how many customers are interested to subscribe the term deposit and after that planning different marketing strategies for them or providing direct marketing efforts to them.

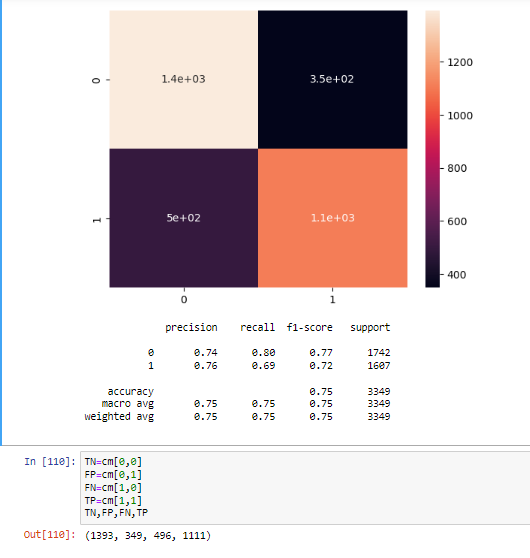
From the above data, we can visualize that it is not a regression problem it’s a classification problem because the output variable is categorical, not numeric, and even it is not continuous so here we are using different classification models. important thing is to feature selection for input and output variables that are decided based on correlation. And after some trial and error, we decided to take the input and output variables as follows:

For output data, there are cases like yes and no means either the customer will be ready to subscribe to the deposit or not ready to deposit for that output variable is categorical so we used the logistic regression. In logistic regression we take the output variable as a categorical variable from that we will get the classification answer.

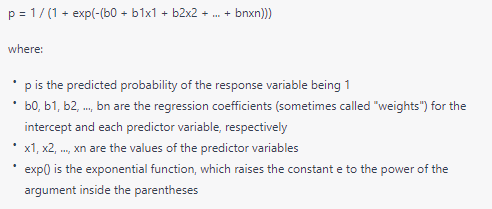
With **logistic regression**, we get 74.39% accuracy these models are trained with the help of the sci-kit learn train\_test\_split model. Like these, we performed all operations to all models like we calculated the confusion matrix for them that shows the TP, TN, FP, and FN values. Also with the sci-kit learn metrics library we plot classification report that contains precision and recall values as well as model accuracy score value. Where the formula for precision and recall are:







Here we get an F1 score is 0.77 and 0.72 for no and yes class resp. here label 0 is no and 1 is yes. the model of logistic regression is as follows:



For every model, we calculated the accuracy score, precision, and recall value and also plot the confusion matrix with the help of the seaborn heatmap.

**Random forest:**

This model is a combination of decision trees it is one of the classification models this model works on an ensemble learning method creating a large amount of decision trees and on that basis they classify and get the accurate output the algorithm’s steps are first of all we randomly selected the training data with the help of sci-kit learns the parameter Random\_state for these we have given 32 random states to train each decision tree then splits the decision tree randomly and the selected subset of the predicted variable then max depth will be decided the model and grow without pruning. After that, we can make predictions on new classification problems and predict the class. random forest models' key benefit is their high accuracy and ability to handle a large number of predictor variables, even when those variables are closely connected. The fact that the individual decision trees are built deep enough to match the training data but not too deep to prevent overfitting makes them resistant to that as well. Here we used random forest and we get an accuracy score of 84.025%. and also we calculated the precision value and recall value with the help of the sklearn module.

**Decision Tree:**

After that, we used a decision tree model that is also a classification model their structure is like a tree where root nodes and leaf nodes are available.

We get a decision trees accuracy is 78.053%

The algorithm behind the model building of the decision tree is they start with the root node for the internal node algorithm selects the feature and splits the data based on purity, it is like the Gini index and entropy. The main important work of the Gini index is to decrease the impurity from the root node to the leaf node. for splitting it to see the highest purity of gain after that it partitioned the dataset into a subset and start corresponding making branches. This algorithm runs till the leaf nodes the of tree get pure. for this model, we get precision and recall value is same that is 0.79 for the no class and 0.77 here also both are the same for the yes type class.

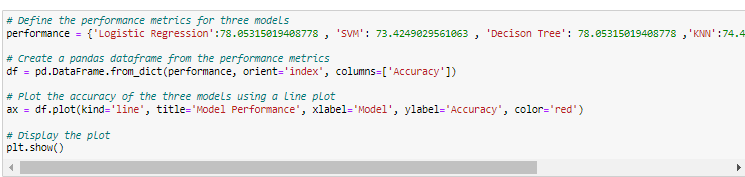
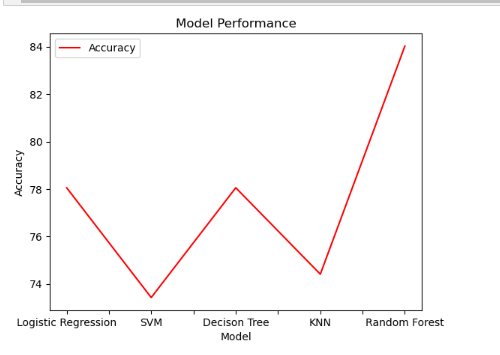
**Support vector Machine (SVM):**

Svm is one of the classification models which split the group into two parts and after that passes one hyperplane to classify the models on that plane basis classes will be distributed. For these model, we get an accuracy value. Of 42% also calculated precision and recall values also this classification is useful for regression it is one of the supervised algorithms here we used this algorithm for the classification of yes or no responses based on the hyperplane. Here we can add parameters like kernel that are useful for converting low dimensional space to high dimensional input space and separable the problem main advantage of this algorithm is it even useful for high dimension cases where kernels present like RBF, poly, linear, etc.

**KNN classifier:**

KNN classifier is one of the classification and regression algorithms here we used that algorithm for classification purposes and it works like when a new, unlabeled data point is found, the algorithm locates the k nearest identified data points and assigns the majority of those neighbors’ labels to the new data point. This algorithm is also called as lazy learner algorithm. For the KNN classifier, we get an accuracy is 74.41%. This algorithm works as first we need to select k neighbors then with the help of Euclidean distance they calculate Euclidean distance and assign the neighbors with the help of k neighbors assigns the data points in each of the categories.

Like these, we get model to output when we compare all these models we get the best accuracy by random forest where accuracy is 84.025% accuracy. After, that we get more accuracy by decision tree which is 75.03%.

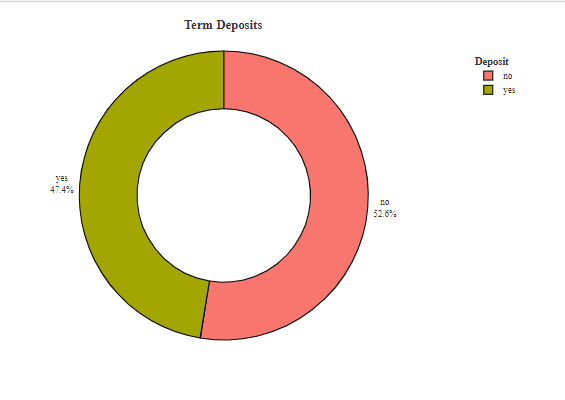


From this code, we get a model performance line plot that helps to easily understand which model gives more accuracy.

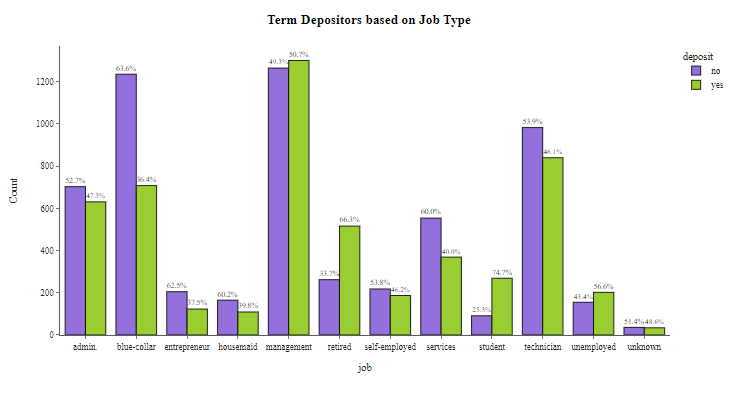
**Visualization:**

From these, we can visualize that 5289 customers are ready to subscribe to the deposit which is 47.4% of the data, and the remaining means 52.6% of customers are not ready for that deal.

Their Dougnut bar structure we have done using Plotly function their visualization is as follows



And also on job type classification we draw the bar plot of term depositors as follows:



**Conclusion:**

From the data we can conclude that most of the customers in the data are lying between the age of 30-40 years, Customers who are married tend to be more likely to open a term deposit account. Also, customers who have higher balances they also more likely to open a term deposit account. Another thing is that most of the customer’s contact was done through the cellular phone. Customers who were previously contacted for a campaign were more likely to open a term deposit account. The majority of the customers who opened a term deposit account did so in the first 15 days of the campaign. And the random forest model accuracy is more than others it is 86% as you saw above so this model is very suitable for these data the bank marketing dataset provides valuable insights into the factors that influence a client's decision to open a term deposit account. The findings can be used to improve future marketing campaigns and to better target potential customers and also useful for making some planes. Some important insights come from the data that help decide on bank marketing.

**References:**

* <https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing>
* <https://www.psycopg.org/docs/>
* <https://www.geeksforgeeks.org/introduction-to-seaborn-python/>
* <https://plotly.com/python-api-reference/>