

PROJECT REPORT



27

Conclusion and Future Steps ..............................................................................................................................

25

Evaluation Metrics ..........................................................................................................................

23

Base Model Building ..............................................................................................................................

21

Class Imbalance ...............................................................................................................................

7

Exploratory Data Analysis ............................................................................................................................

7

Project Justification ............................................................................................................................

6

Unknown Values .............................................................................................................................

18

Statistical Significance of variables ...................ofof...........................................................................................................................................

5

count of missing values ...............................................................................................................................

5

Preprocessing Data Analysis.....................................................................................................................................

3

Dataset Description DEDe.DescriptionDescription................................................................................................................................ escrip.................................................................................................................................

3

Literature Survey ........................................................................................................................

3

Industry Review ............R..ReviewReview...................................................................................................................................

**CONTENTS**

**Industry Review**

In the banking sector, direct marketing campaigns are a common practice used to promote products such as term deposits, loans, and credit cards. These campaigns typically involve contacting potential customers through various channels, such as phone calls, emails, and direct mail. The goal is to identify and convert leads into customers, thereby increasing the bank’s revenue and customer base.

**Literature Survey**

Numerous studies and publications focus on the application of machine learning and statistical models to improve the effectiveness of marketing campaigns. Previous research has explored various algorithms for customer segmentation, predictive modeling, and response optimization. Key publications in this area include studies on logistic regression for response prediction, the use of decision trees and random forests for classification tasks, and the application of boosting algorithms like XGBoost for improving model accuracy.

**Dataset Description:**

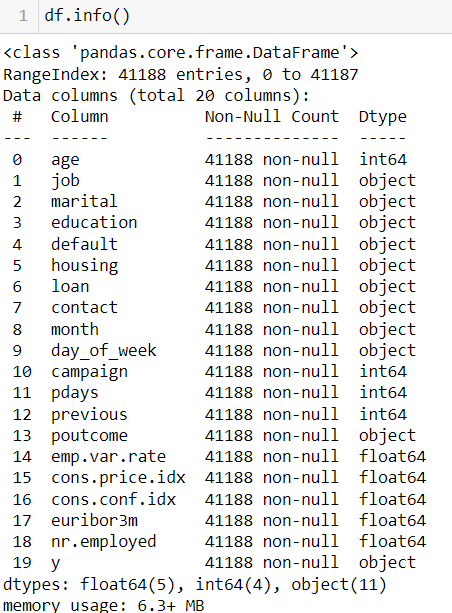
**Data Dictionary:**

Here is the description of the dataset:

|  |  |
| --- | --- |
| Column Name | Description |
| Age  (numeric) | Age of client |
| job  (categorical) | type of job ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown') |
| Marital  (categorical) | marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed) |
| Education  (categorical) | ' Basic.4y','basic.6y','basic.9y','high. school, 'illiterate', 'professional. course', 'university. Degree', 'unknown' |
| Default (categorical) | has credit in default? (categorical: 'no', 'yes', 'unknown') |
| Housing (categorical) | has housing loan? ('no', 'yes', 'unknown') |
| Loan  (categorical) | has personal loan? ('no', 'yes', 'unknown') |
| Contact  (categorical) | Communication type (‘cellular’, ‘telephone’) |
| Month  (categorical) | Last contact month of the year (‘jan’ to ‘dec’) |
| Day\_of\_week  (Categorical) | Last contact day of the week (‘mon’ to ‘fri’) |
| Campaign  (numeric) | number of contacts performed during this campaign and for this client |
| Pdays  (numeric) | number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) |
| Previous  (numeric) | number of contacts performed before this campaign and for this client |
| Poutcome  (categorical) | outcome of the previous marketing campaign ('failure', 'nonexistent’, ‘success') |
| emp.var.rate  (numeric) | employment variation rate - quarterly indicator |
| Cons.price.idx  (numeric) | consumer price index - monthly indicator |
| Cons.conf.idx  (numeric) | consumer confidence index - monthly indicator |
| Euribor3m  (numeric) | euribor 3-month rate - daily indicator |
| Nr.employed  (numeric) | number of employees - quarterly indicator |
| **Output Variable (Target)** | |
| Y  (categorical) | has the client subscribed a term deposit? (binary: 'yes', 'no') |

**Variable categorization (count of numeric and categorical)**

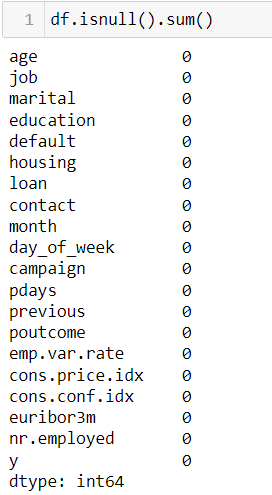
11 categorical variables and 9 numerical variables (5 continuous-numeric and 4 discrete-numeric)



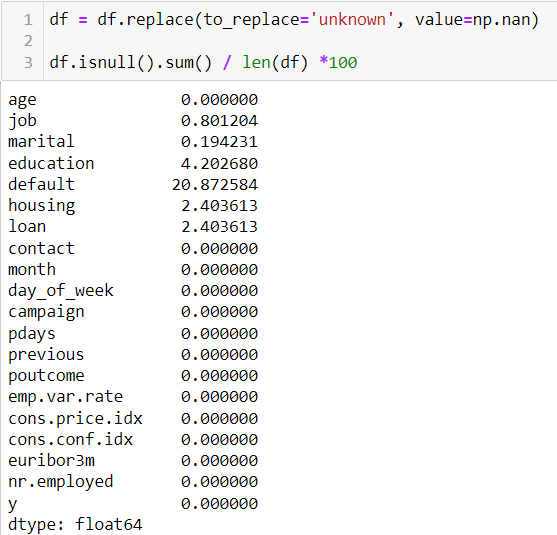
**Preprocessing Data Analysis:**

Pre-Processing Data Analysis (count of missing/ null values, redundant columns, etc.)

**count of missing values** - Initially, we had a no missing value



**Unknown values**:

We have seen that there are no null values in our dataset. But the search does not stop here as we've only cross-verified for our numeric columns. For the categorical columns, there are high chances of null values to be equivalent to 'unknown' so we shall convert 'unknown' to NA to finally count the total number of missing values in our categoric columns 

**Alternate sources of data that can supplement the core dataset (at least 2-3 columns)**

The data is enriched by the addition of five new social and economic features/attributes (national wide indicators from a ~10M population country), published by the Banco de Portugal and publicly available at: <https://www.bportugal.pt/estatisticasweb>.

This dataset is almost identical to the one used in [Moro et al., 2014] (it does not include all attributes due to privacy concerns). Using the rminer package and R tool (<http://cran.r-project.org/web/packages/rminer/)>, we found that the addition of the five new social and economic attributes (made available here) lead to substantial improvement in the prediction of a success, even when the duration of the call is not included. Note: the file can be read in R using: d=read.table("bank-additional-full.csv",header=TRUE,sep=";")

**Project Justification:**

**Problem** **Objective**: The classification goal is to predict if the client will subscribe (yes/no) to the term deposit

**Complexity Involved**: Handling imbalanced data, feature selection, feature engineering, model building and evaluation.

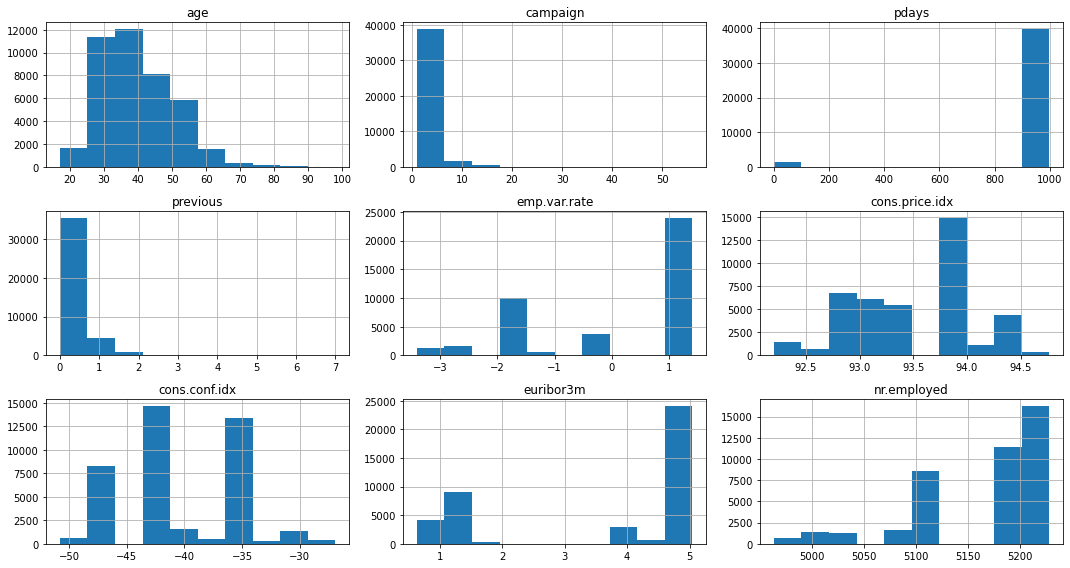
**Project Outcome**: The outcome will have commercial value by improving campaign effectiveness and customer conversion rates. Increases bank revenue by identifying potential subscribers, thus enabling focused marketing efforts. Utilize machine learning to analyse historical data, leading to data-driven decision-making.

**Exploratory Data Analysis:**

**Univariate Analysis: (numerical variables)**

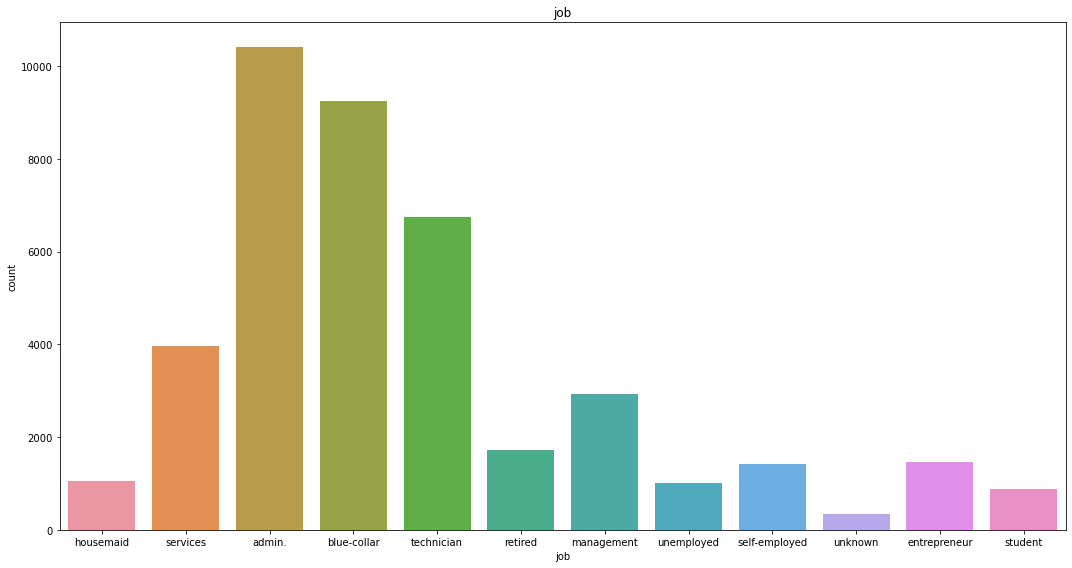
**Distribution of numerical variables**

We used histplot, for finding the distribution of numerical variables

****

**Univariate Analysis: (Categorical variables)**

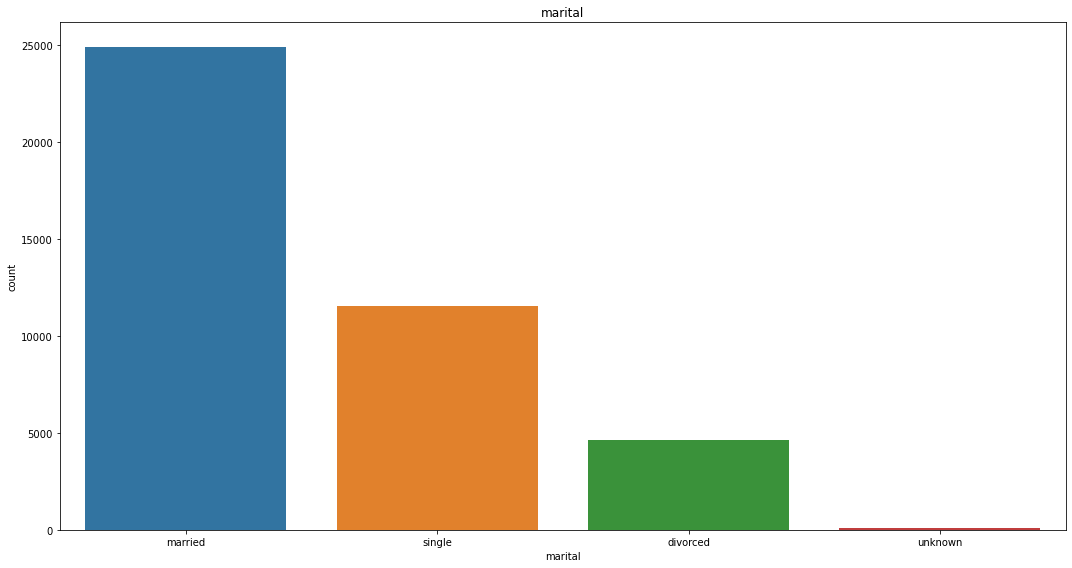
**Job:**

****

**Inferences**:

* Here Job Types, 'Blue-collar' and 'management' professions are the most common among clients
* Tailoring campaigns to the needs and interests of these job types can increase engagement and response rates

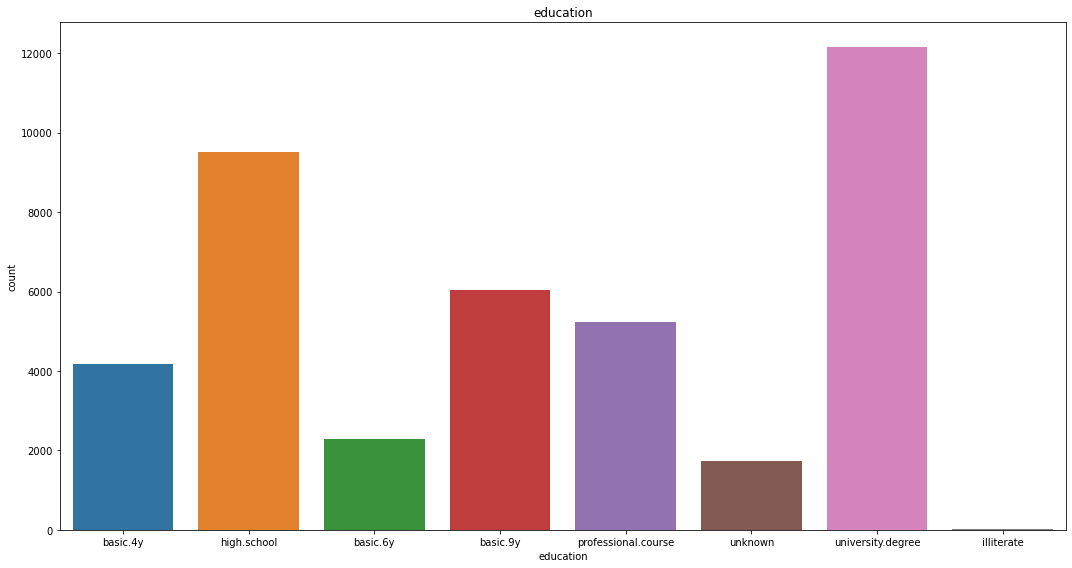
**Marital:**

****

**Inferences**:

* Married clients dominate the dataset
* Designing campaigns that cater to married individuals, such as family-oriented products or services, could enhance effectiveness and appeal

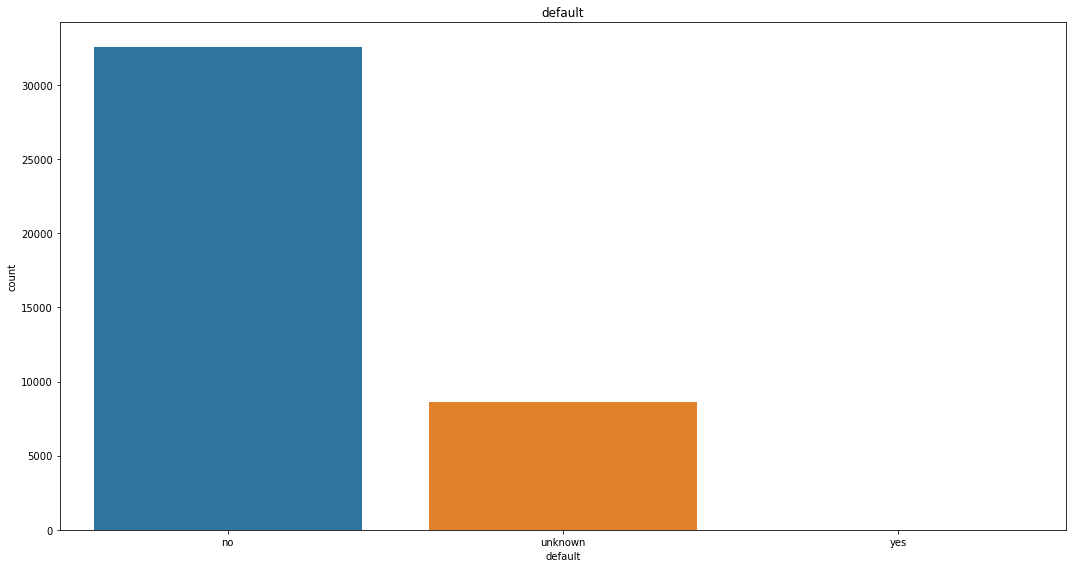
**Education:**



**Inferences:**

* Most clients have a 'university. Degree' or 'high. School' education
* Incorporating educational content and emphasizing product benefits aligned with their background can resonate well with these clients

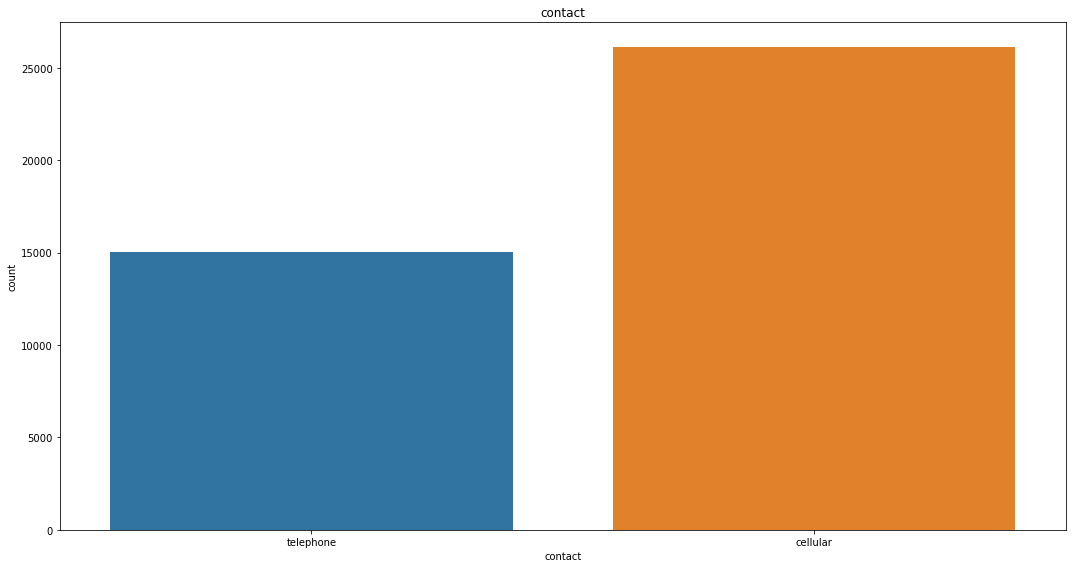
**Default:**

****

**Inferences:**

* The vast majority of clients do not have credit defaults, indicating financial stability
* Marketing strategies should focus on investment opportunities or premium offerings appealing to financially stable individuals

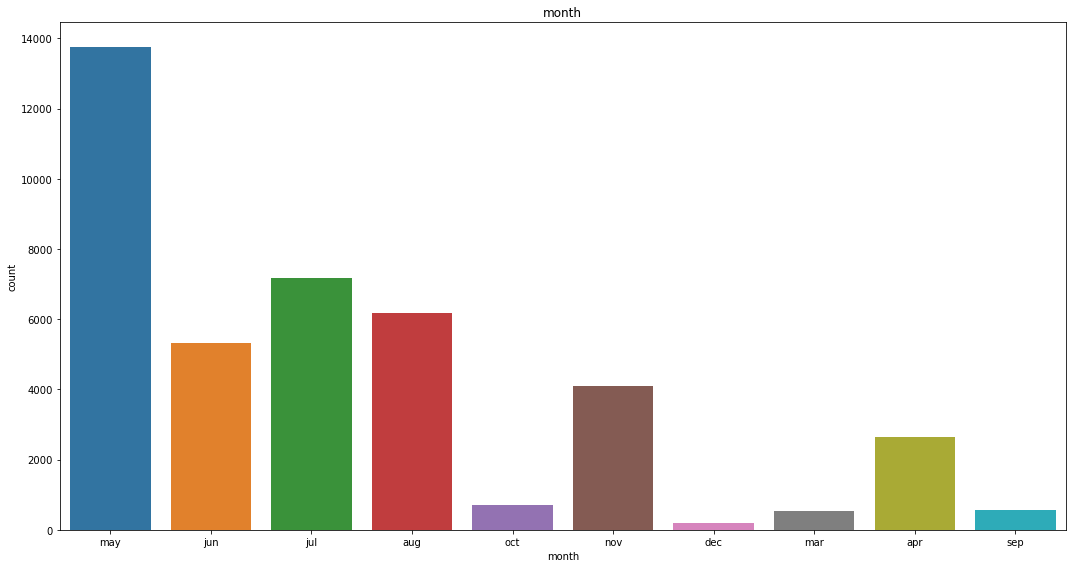
**Contact mode:**

****

**Inferences:**

* The 'cellular' contact method is more prevalent and effective than 'telephone' contact
* Prioritizing mobile outreach can enhance engagement and conversion rates

**Month:**

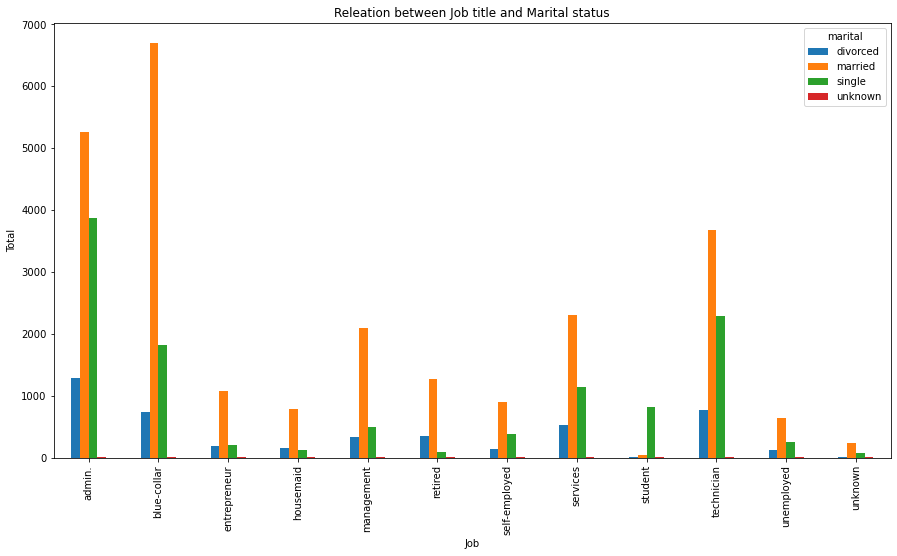
****

**Inferences:**

* May has the highest number of contacts, making it strategic for campaign launches
* Concentrating marketing efforts in May can capitalize on increased customer responsiveness

**Bivariate Analysis: (cat vs cat)**

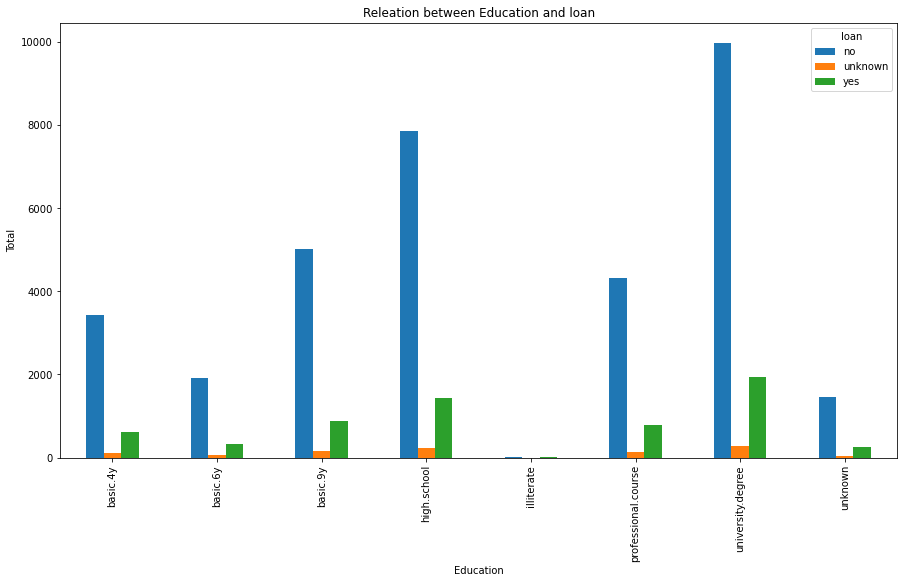
**Job vs Marital:**

****

**Inferences:**

* Married clients are prevalent across most job categories, especially 'blue-collar' and 'management'
* Marketing strategies should consider the dual-income and family-oriented needs of married clients

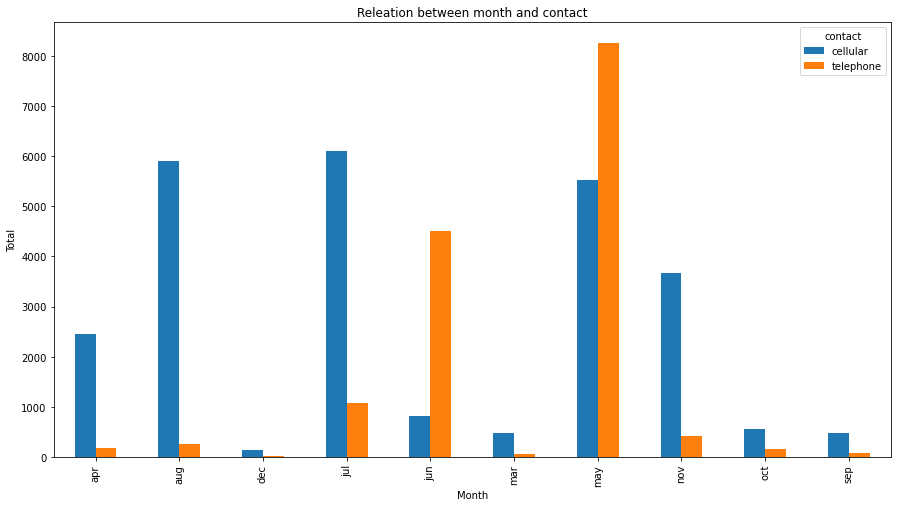
**Education vs loan:**

****

**Inferences:**

* Higher education levels correlate with fewer personal loans, suggesting financially prudent behaviors
* Financial products emphasizing savings and investments might appeal to educated clients

**Month vs contact method:**

****

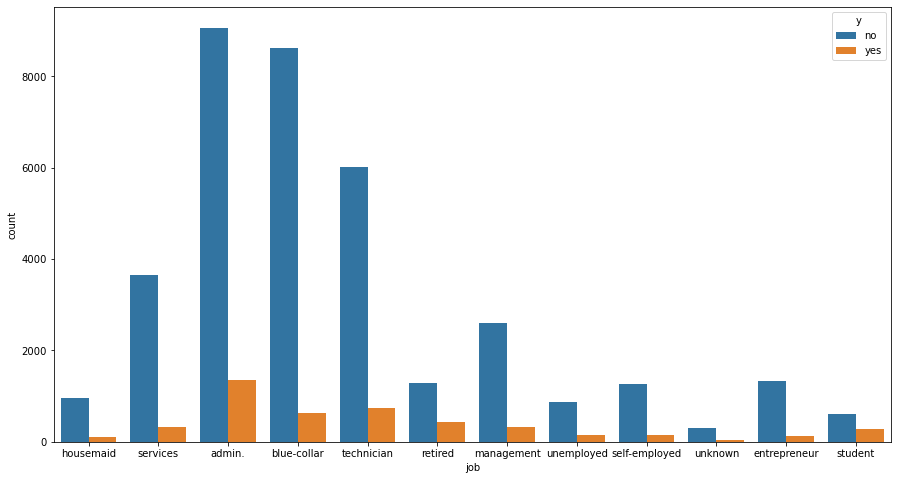
**Inferences:**

* 'Cellular' contact is more effective than 'telephone' across all months
* Prioritizing mobile outreach can improve campaign reach and effectiveness

**Analysis with Target variable:**

We have analysed some of the features with target variable. We got some efficient insights from that they are listed below:

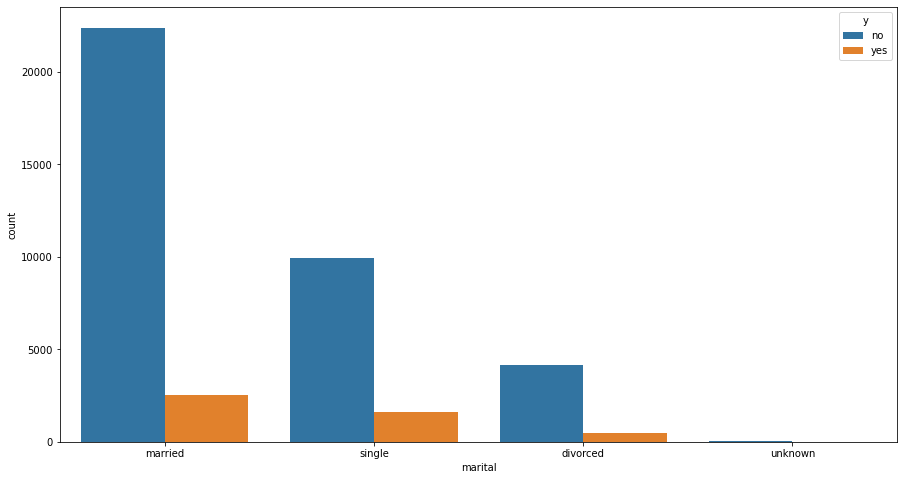
**Job vs Target variable:**

****

**Inferences:**

* Clients in 'admin.' and 'blue-collar' jobs show higher 'no' responses, while 'admin' and 'technicians' clients have more 'yes' responses
* Tailoring messages to address the specific concerns of 'admin.' and 'technicians' workers, while highlighting benefits relevant to 'blue-collar' and 'retired' clients, can optimize campaign success

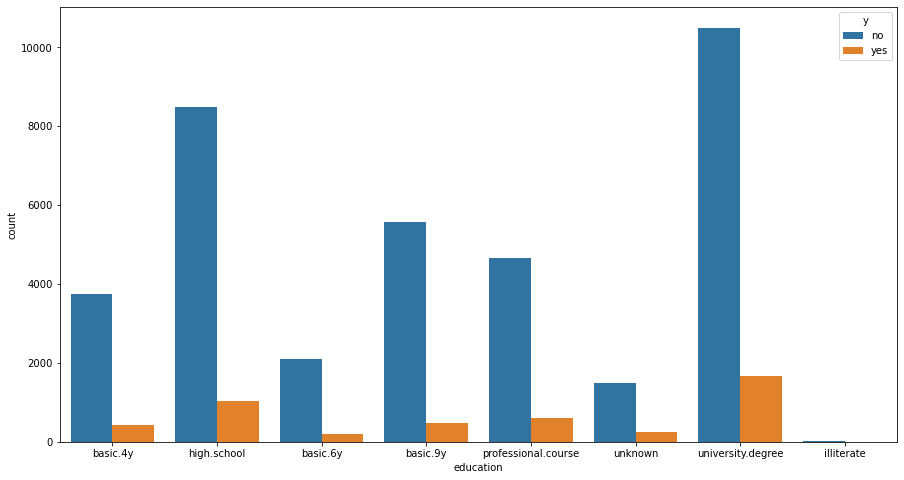
**Marital status vs Target variable:**

****

**Inferences:**

* Married clients have a higher number of 'no' responses, whereas single clients have a relatively higher proportion of 'yes' responses
* Campaigns designed to appeal to single clients’ needs and lifestyles might achieve better results

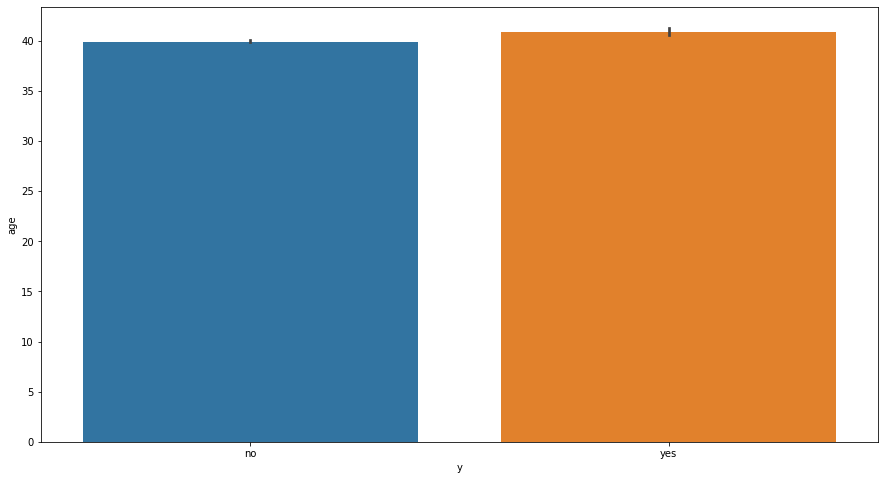
**Education vs Target:**

****

**Inferences:**

* Clients with a 'university.degree' show more 'yes' responses compared to those with only 'basic.4y' education
* Marketing efforts should emphasize educational and professional development benefits to attract this more responsive segment

**Age vs Target:**

****

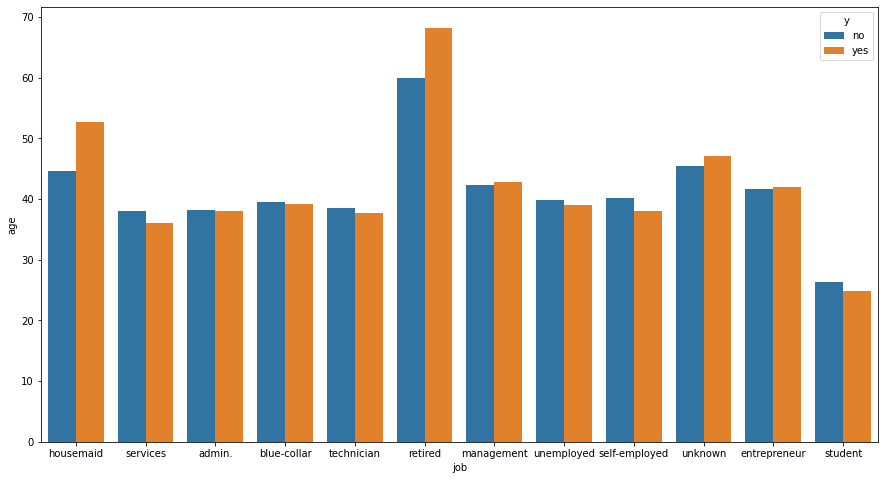
**Inferences:**

* Younger clients are less likely to subscribe to term deposits compared to older clients
* Targeting marketing strategies towards older demographics, who show higher interest, can increase subscription rates

**Multivariate analysis:**

Basically, multivariate analysis is analysing more than 2 variables. We have done multivariate analysis with some of the features with target variable. They are:

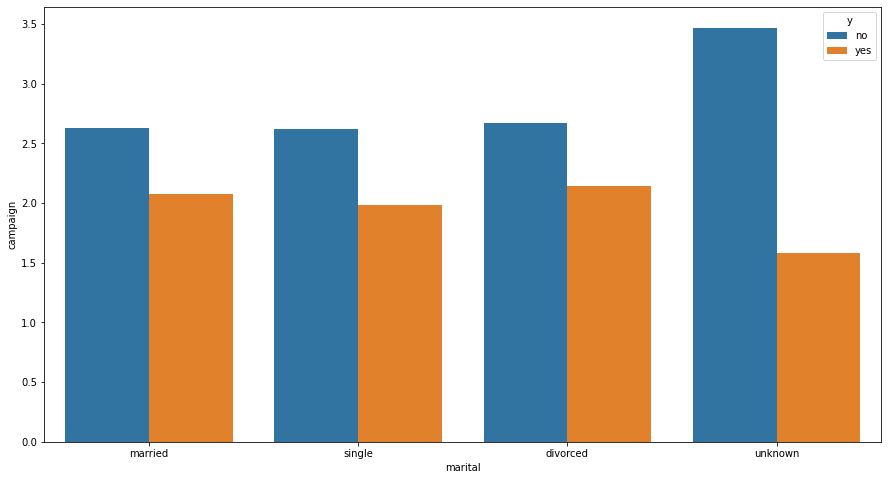
**Job vs Age vs Target:**

****

**Inferences:**

* 'Retired' and 'housemaid' clients, who tend to be older, have higher positive response rates
* Focusing on these segments can increase campaign success

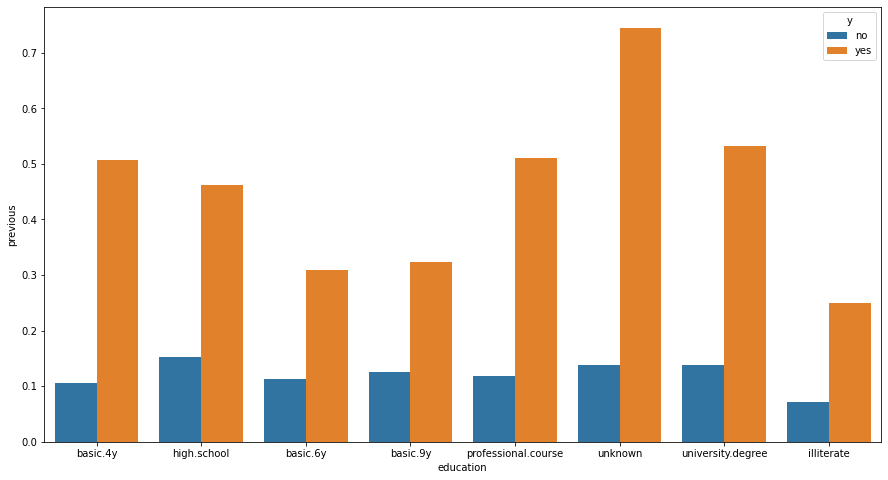
**Marital status vs campaign vs Target:**

****

**Inferences:**

* Single clients require fewer contacts for a positive response
* Strategies targeting single clients could be more cost-effective

**Education vs previous contact vs Target:**

****

**Inferences:**

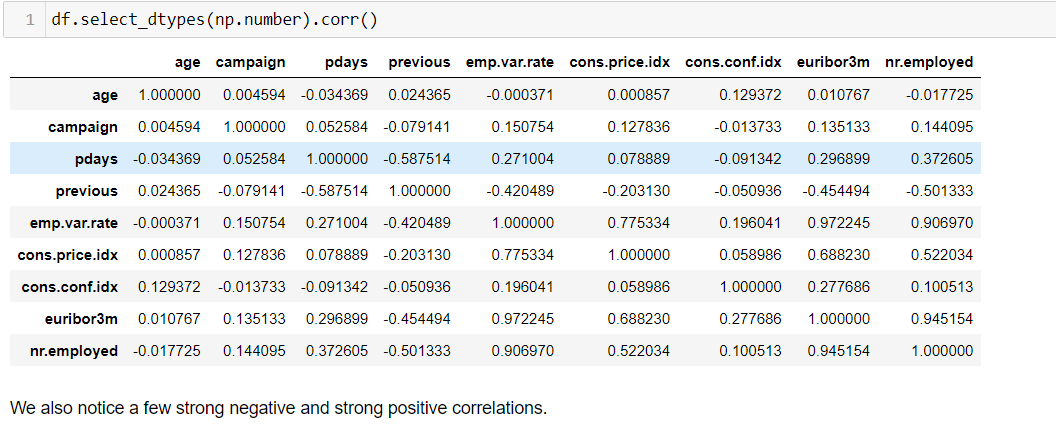
* Clients with 'high school' show higher positive response rates and more previous contacts
* Prior successful engagement strategies should be replicated for educated clients

**Correlation Matrix:**

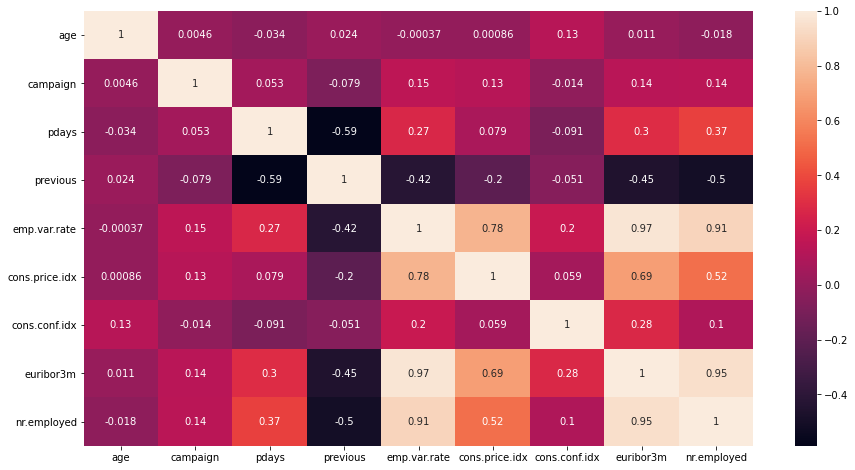
A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. The value ranges between -1 and 1.

* 1 implies a perfect positive correlation
* -1 implies a perfect negative correlation
* 0 implies no correlation

We got correlation matrix from that we noticed some of the variables are strongly correlated in both negative and positive directions.



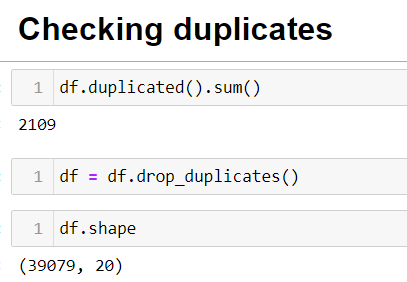
**Graphical Representation:** We used heatmap for better understanding how variables are correlated.



**Identification of Duplicates:**

Duplicate records can cause a model to overfit, meaning it performs very well on the training data but poorly on new, unseen data. The model might learn the noise or redundant patterns instead of the underlying trends. In statistical models, duplicates can skew the estimates of parameters, leading to biased or incorrect conclusions. Duplicates can inflate performance metrics like accuracy, precision, and recall, giving a false impression of the model’s effectiveness. It can increase the computational complexity of the model training process, making it slower and more resource-intensive. Duplicates can affect the correlation matrix, as they artificially inflate the correlation coefficients, leading to misinterpretation of the relationships between variables.

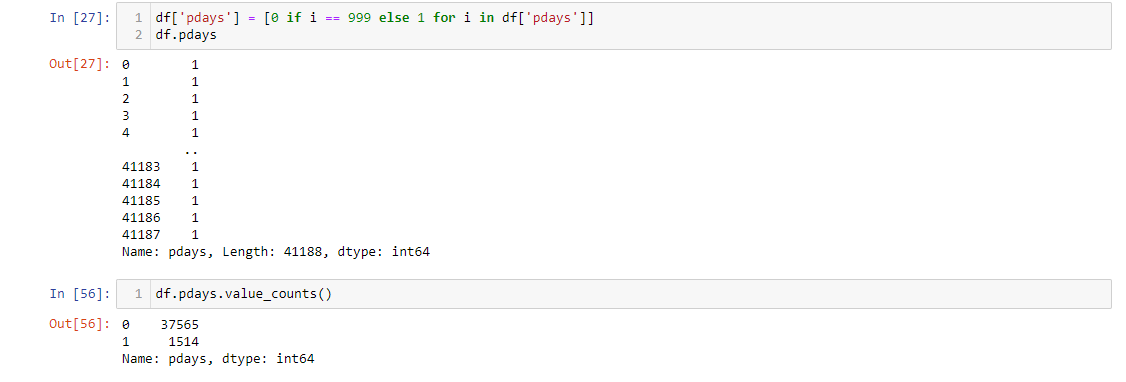
**Treatments:** We removed duplicates from the dataset since it will affect our model.

****

**Note:** After removing duplicates, we removed around 2000 record

**Pdays Modification :**

Here We have numeric attribute named pdays means that number of days that passed by after the client was last contacted from a previous campaign (There are 999 and other numeric values inside of this column. 999 means client was not previously contacted. We converted the ‘pdays’ column numeric to categorical which as two classes. If the value is equal to 999, we put the 0 instead of 999, otherwise we put 1. So we changed the pdays column.

****

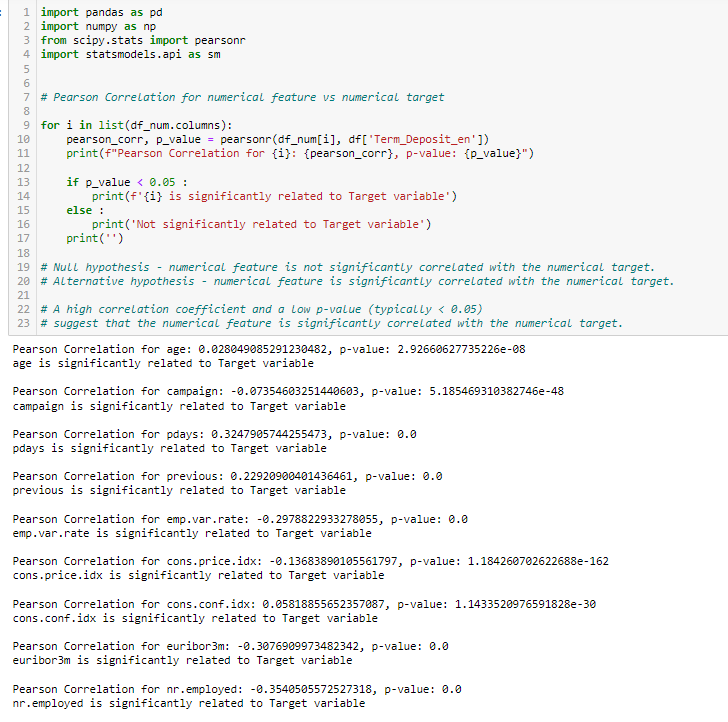
**Statistical significance of variables:**

Assessing the statistical significance of variables in a dataset is a crucial step in understanding which variables have a meaningful impact on the target variable. Statistical significance is typically determined through hypothesis testing, which helps you understand whether the observed relationships in your data occurred by chance or reflect actual relationships in the population.

**Pearson Correlation coefficient:**

The Pearson correlation coefficient is a measure of the linear relationship between two variables. It ranges from -1 to 1, where:

* 1 indicates a perfect positive linear relationship,
* -1 indicates a perfect negative linear relationship,
* 0 indicates no linear relationship.



We have done Pearson correlation test which is having null and alternative hypothesis as:

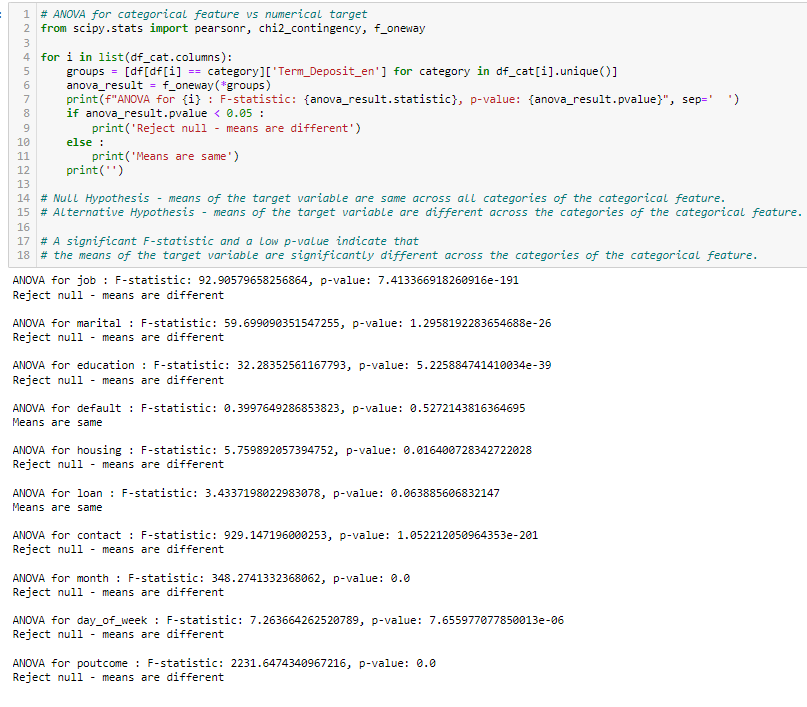
Null hypothesis - numerical feature is not significantly correlated with the numerical target.

Alternative hypothesis - numerical feature is significantly correlated with the numerical target.

From this test some of the features are significantly related to target variable they are age, campaign, pdays etc.

**ANOVA: Analysis of Variance**

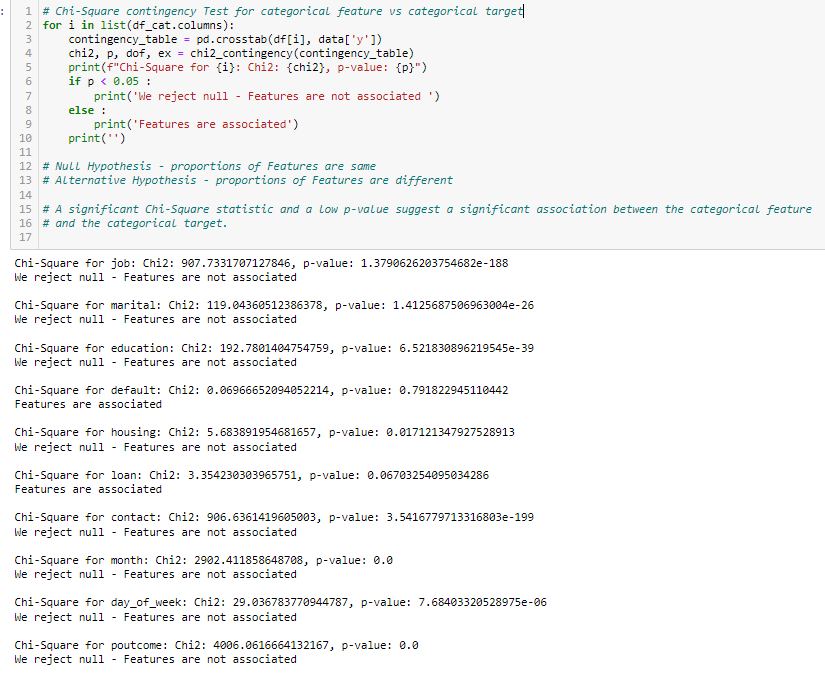
Analysis of Variance (ANOVA) is a statistical method used to test the differences between the means of three or more groups. It helps to determine if at least one of the group means is statistically different from the others. ANOVA is useful when comparing multiple groups to see if they have different effects on a dependent variable.



We have done anova for numerical variable vs categorical variable with more than 2 classes. And we found some of the categories with more than 2 classes are having different means.

**Chi- Square Test for Independence:**

The Chi-Square Test for Independence is a statistical test used to determine if there is a significant association between two categorical variables. It tests the null hypothesis that the variables are independent.



The Chi-Square Test for Independence is a useful tool for determining if there is a significant relationship between two categorical variables. By comparing observed and expected frequencies, you can infer whether the variables are associated or independent. This test is widely used in fields like marketing, research, and social sciences to analyze survey data and experimental results. We found that some of the independent features which are categorical are not associated that much.

**Inferences from Statistical Test :** After performing ANOVA and Chi-Square Contingency test we found features **Loan and Default** as non significant . As a result we will be building models without these features to improve the accuracy and F1 score .

**Class Imbalance:**

Class imbalance occurs when the classes in a classification problem are not represented equally. This can be problematic because machine learning models might become biased towards the majority class, leading to poor performance on the minority class.

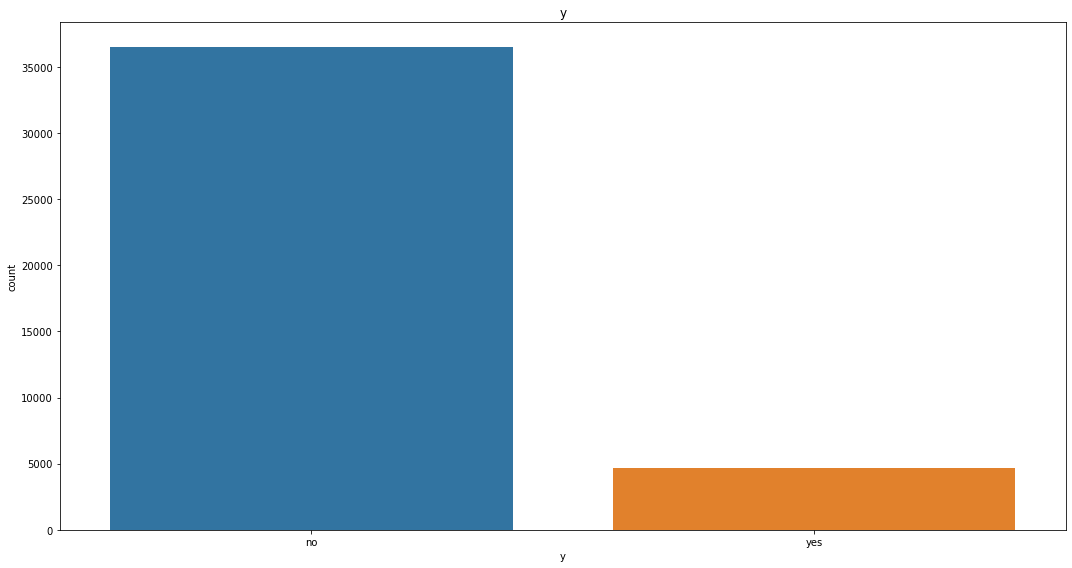
**Issues Caused by Class Imbalance:**

**Bias Towards Majority Class:** The model may perform well on the majority class but poorly on the minority class.

**Misleading Accuracy:** High accuracy might be misleading as the model could simply be predicting the majority class most of the time.

**Poor Generalization:** The model might not generalize well to unseen data, particularly for the minority class.

From our data we found that target variable is strongly imbalanced. No class is majority and yes class is minority.



**Methods to Handle Class Imbalance:**

**Resampling Techniques:**

**Oversampling the Minority Class:** Increase the number of instances in the minority class by duplicating them.

**Under sampling the Majority Class**: Reduce the number of instances in the majority class by randomly removing them.

**Synthetic Data Generation (SMOTE):** Create synthetic instances of the minority class.

**Base Model Building**

**Encoding:**

Encoding is the process of converting categorical variables into numerical values that can be used by machine learning algorithms. Different encoding techniques are used depending on the nature of the categorical data and the requirements of the model.

**Common Encoding Techniques**

**Label Encoding**

Label Encoding assigns a unique integer to each category. This method is straightforward but can introduce an ordinal relationship where none exists.

**One-Hot Encoding**

One-Hot Encoding creates a new binary column for each category, which has a value of 1 for the presence of the category and 0 otherwise. This method avoids the ordinal relationship issue.

**Ordinal Encoding**

Ordinal Encoding assigns ordinal values to categories based on a specific order. This is useful when the categories have a meaningful ranking.

**Frequency Encoding**

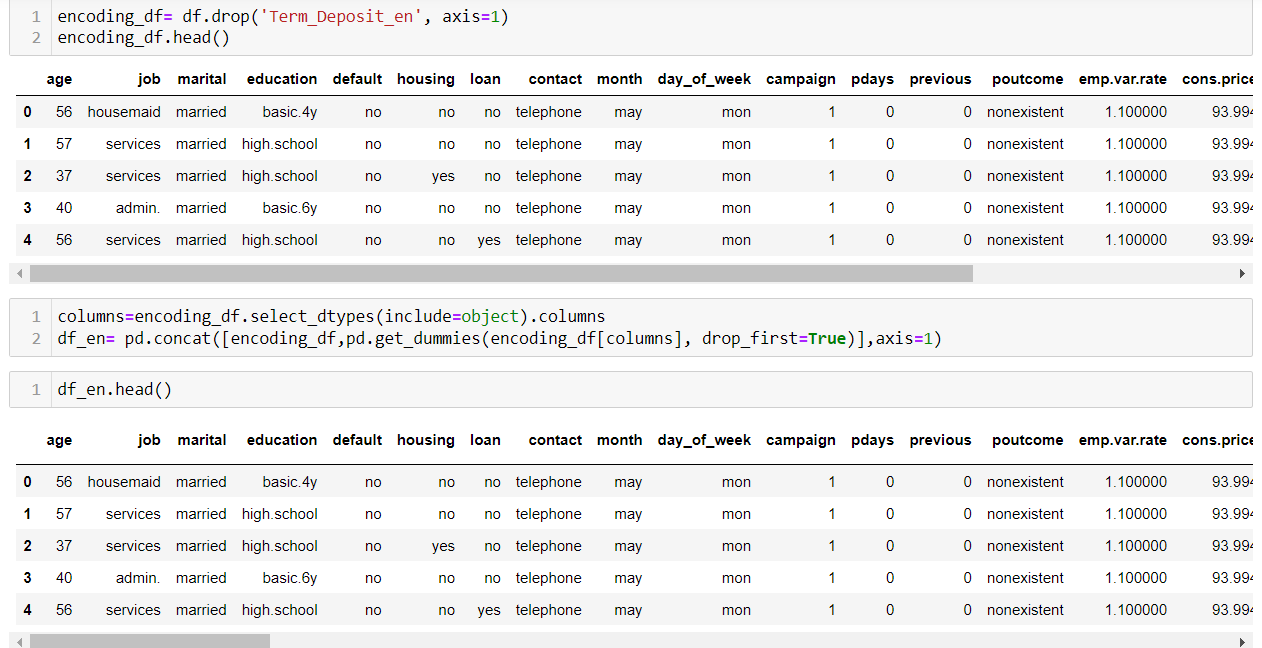
Frequency Encoding replaces categories with their frequency of occurrence. This method can be useful for high cardinality features.

**Target Encoding**

Target Encoding replaces categories with the mean of the target variable for each category. This is often used in categorical features with high cardinality.

Choosing the right encoding technique is crucial for the performance of machine learning models. While label encoding is simple and useful for ordinal data, one-hot encoding is generally preferred for nominal data. More advanced techniques like binary encoding, ordinal encoding, frequency encoding, and target encoding can be advantageous depending on the dataset and the specific problem. Proper encoding ensures that categorical variables are effectively utilized, leading to better model accuracy and performance.

**We have used one-hot encoding technique for our categorical variables.**



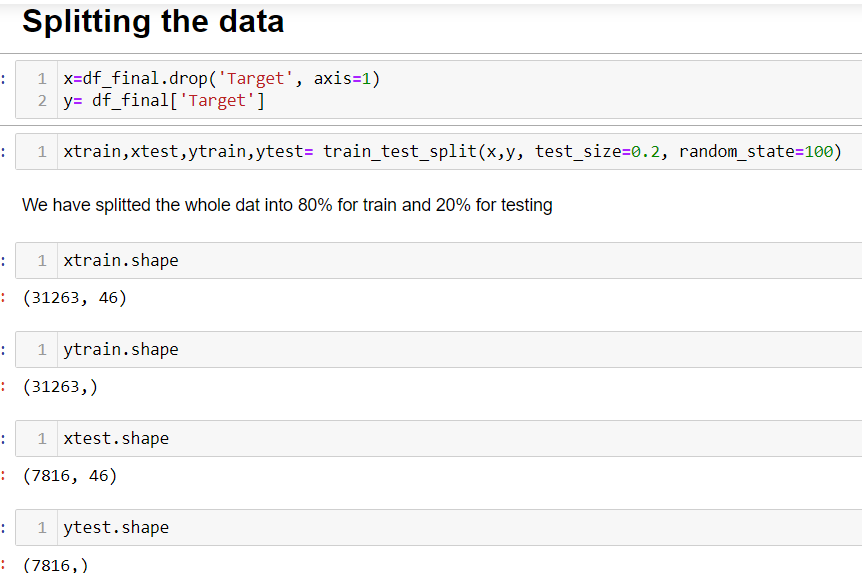
**Splitting the data**:

Splitting data is a fundamental step in machine learning where you divide your dataset into training and testing sets. This allows you to train your model on one subset (training set) and evaluate its performance on another subset (testing set). Here's how you can split your data using Python:

Splitting Data Using train\_test\_split from sklearn.model\_selection

The train\_test\_split function from sklearn.model\_selection is commonly used for splitting datasets. It shuffles the data and splits it into two or more parts according to the specified proportions.

Here we divided the data into 2 parts which training parts is 80% of data and testing part is 20% of data.



Splitting your data correctly ensures that your model is trained on a diverse set of data and evaluated on unseen data, which helps in assessing its generalization performance.

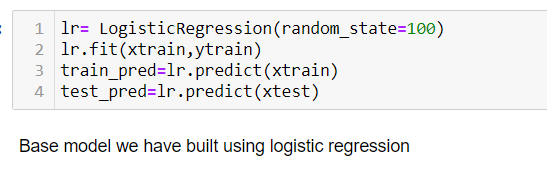
As we are using all classification techniques, KNN model is a distance based model for which scaling is required. So we have also scaled the data and then again performed train test split.

**Base Model Algorithm:**

Based on the problem statement and our dataset it’s classification problem. We built base model using logistic regression algorithm.

**Logistic Regression:**

Logistic regression is a statistical model used for binary classification tasks, where the target variable (dependent variable) is categorical with two possible outcomes, often coded as 0 and 1. It's a type of regression analysis that estimates the probability of the outcome based on one or more predictor variables (independent variables).

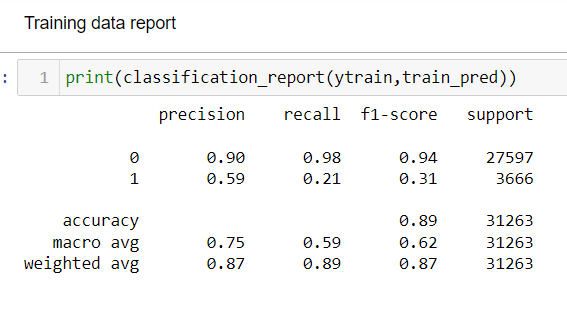


**Evaluation Metric:**

We used Classification Report for both training and testing data. As our target variable is imbalanced, we can’t rely on accuracy so we are relying on f1\_score and also plotted the confusion matrix.

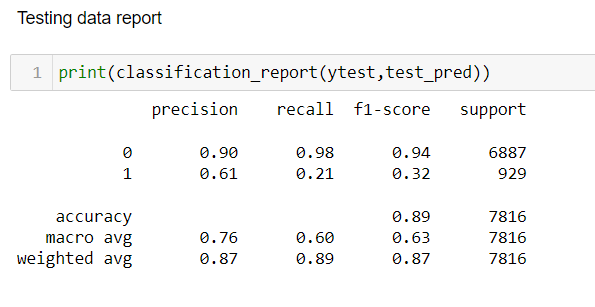
**Training data report:**

From classification report we found out that training accuracy is 0.87



**Testing data report:**

From classification report we found out that testing accuracy is also 0.87

****

**Confusion Matrix:**

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It allows visualization of the performance of an algorithm and provides insights into how well the model is performing in terms of predicting each class.

**Components of a Confusion Matrix:**

In a binary classification problem, the confusion matrix consists of four main metrics:

True Positive (TP): The number of correctly predicted positive instances (actual positive and predicted positive).

True Negative (TN): The number of correctly predicted negative instances (actual negative and predicted negative).

False Positive (FP): Type I error - the number of incorrectly predicted positive instances (actual negative but predicted positive).

False Negative (FN): Type II error - the number of incorrectly predicted negative instances (actual positive but predicted negative).

**Interpretation:**

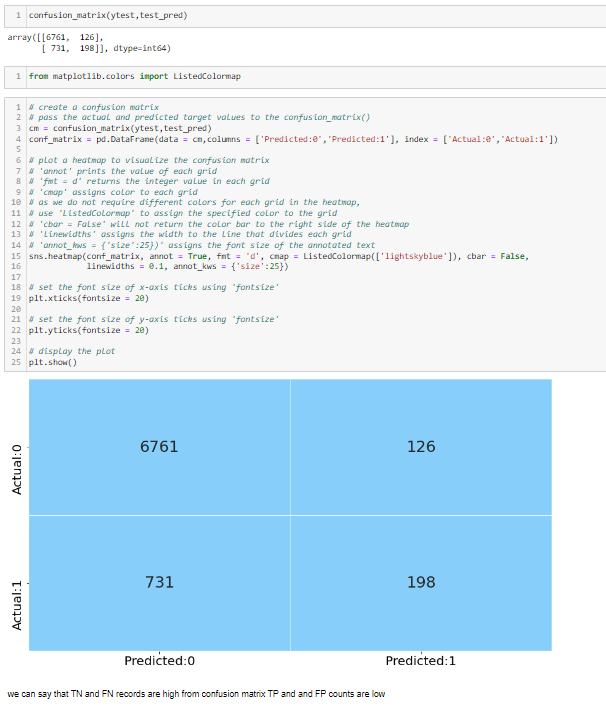
True Positive (TP): The model correctly predicted the positive class.

True Negative (TN): The model correctly predicted the negative class.

False Positive (FP): The model incorrectly predicted the positive class (Type I error).

False Negative (FN): The model incorrectly predicted the negative class (Type II error).

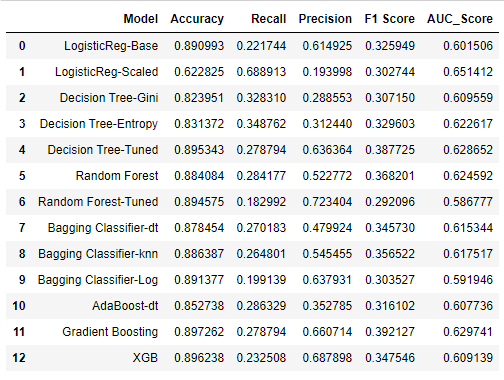
We have calculated number of records in each segment and plotted as matrix. So, we checked the values of confusion matrix to see which records are under True positive, False positive, False negative and True Negative.



**Model Selection :**

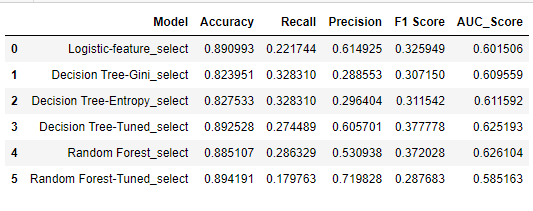
List of Algorithms: Logistic Regression, Decision Trees, Random Forests, Bagging Classifier, Ada-Boost, Gradient Boost, XG Boost.

We performed basic model building from this list of models and also performed their hyperparamter tuning. And finally we got 13 models . The summary of every model is shown below .



After analyzing the accuracy value, f1 score and AUC score of all models we foundtwo models **“Decision Tree Tuned “** and **“Gradient Boosting” model** to have the best values.

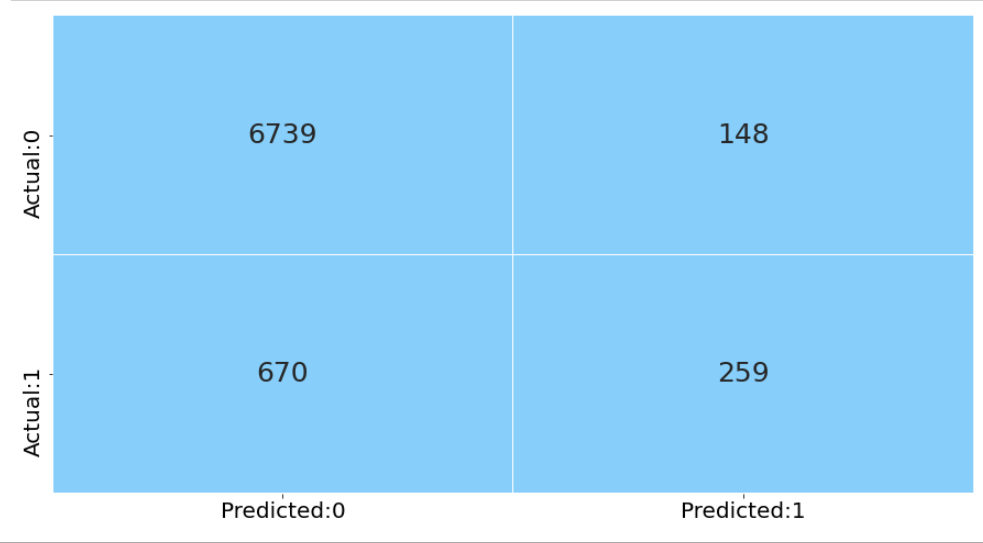
As mentioned, we also performed model building after feature selection i.e, after removing Loan and Default features to check and achieve higher accuracy and F1-score . The table shown below contains the results .

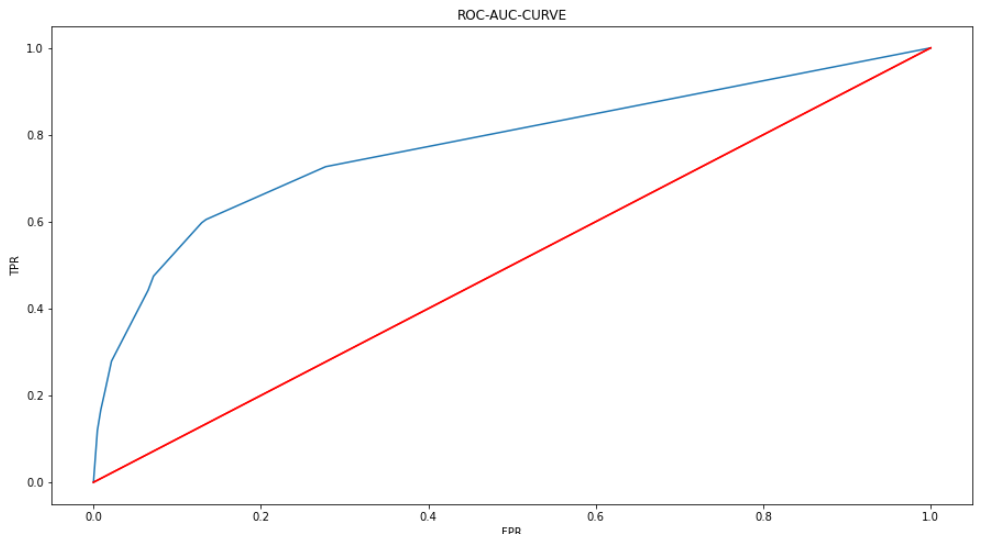


From this approach we can see that there is no significant change in the values of accuracy and F1 Score .

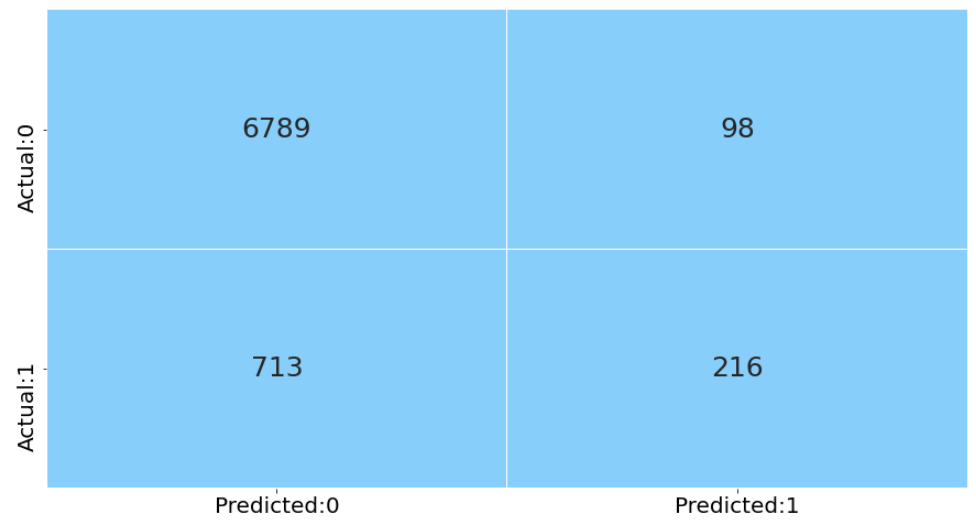
**Here are the plots of confusion matrix and ROC curve for the selected best models.**

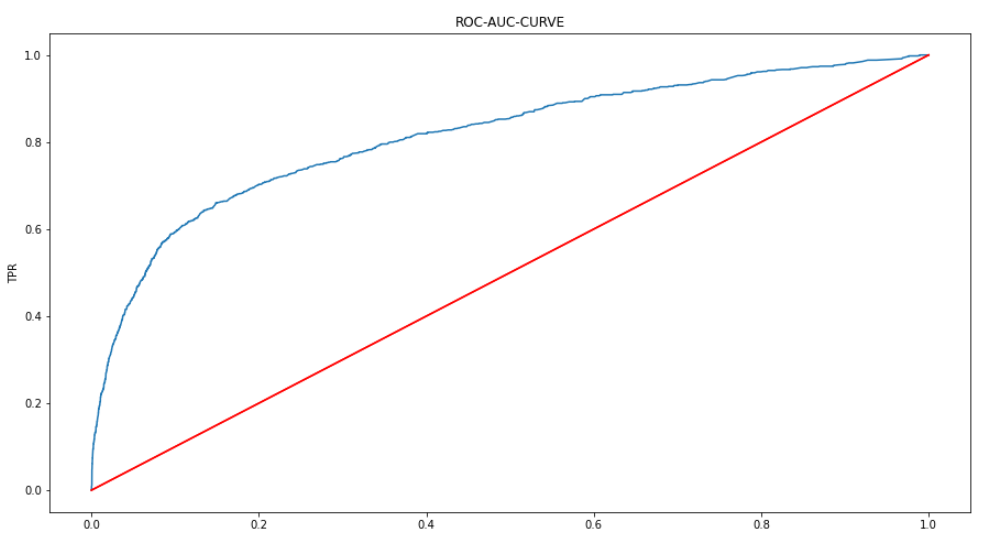
**Decision Tree Tuned**

****

****

**Gradient Boosting**

****

****

**Conclusion and Future Steps:**

* More accuracy can be achieved using more complex models .
* explore additional features, and apply segmentation analysis for better-targeted campaigns.
* Remodelling can be done after SMOTE Technique.
* Perform CV fold techniques to prevent data leakage.