Project Summary

|  |  |
| --- | --- |
| Batch details | DSE FEB 2021 |
| Team members | Bhawyaraj Panicker  Pauline Rathni Grace  Rhea Prakash  Syed Mohd Emad Imam  Vijaydeep Mohapatra |
| Domain of Project | Recommendation System |
| Proposed project title | Bank Campaign Recommendation System |
| Group Number | 5 |
| Team Leader | Syed Mohd Emad Imam |
| Mentor Name | Mr. Animesh Tiwari |

Date: 31/07/2021

Signature of the Mentor - Mr. Animesh Tiwari

Signature of the Team Leader - Syed Mohd Emad Imam

Table of Contents

|  |  |  |
| --- | --- | --- |
| Sl NO | Topic | Page No |
| 1 | Overview | 3 |
| 2 | Business problem goals | 3 |
| 3 | Topic survey in depth | 4 |
| 4 | Critical assessment of topic survey | 5 |
| 5 | Methodology to be followed | 7 |
| 6 | References | 13 |

Project Details

OVERVIEW

Recommendation systems are becoming increasingly important with its accuracy in personalized suggestion based on a pattern found. Nowadays, we see recommendation systems being used almost everywhere in our day to day lives. These recommendations are mostly accurate and precise, enabling us to make better decisions conveniently. Here, we have the bank campaigning recommendation system that helps banks make a decision at ease as to which campaign, they should go for and focus more on the customers who are likely to subscribe. Our primary goal was to understand, analyze and correlate various ML algorithms to direct those customers to appropriate campaigns.

# Summary of problem statement

# 

1. Problem Understanding

The purpose is to build a bank campaign recommendation system to the appropriate customers who are likely to subscribe. By identifying patterns we would like to propose a system that would get this work done. This would help the bank to send an appropriate campaign at ease to the suitable customers which would benefit both the parties. We aim at creating various ML algorithms for this problem and use the one with better accuracy and precision to suggest these recommendations.

2. Objective

The recommender system is useful to any business that makes money via recommendations. We have taken the ‘Portuguese banking institution’ dataset for this purpose. Giving good recommendations will help the banks improve their business and also at the same time deliver to the customers as to exactly what they are looking for. This will help the customer to carry on with the best service and have a good experience.We also believe that recommendation systems play a vital role in the banking platforms, where they invest in data science to improve their customer satisfaction by attracting new subscribers to their plans through proper campaigns. This also helps the bank understand the customer churn better.

1. Approach

We aim at creating a ML model that recommends a particular bank how they can use predictive analytics to help prioritize customers who would subscribe to a bank term deposit. We are using supervised classification learning to predict the model.

1. Conclusions

Based on various ML algorithms and detailed exploratory analysis of all the factors that contribute to this decision, we aim to improve the accuracy and personalize recommendations based on the patterns observed by choosing a more accurate ML model for classification problem.

# Overview of the final process

* Our targeted system was the Bank campaigning recommendation system and our goal was to understand , analysis and correlate the various ML algorithms to direct the targeted audience who would subscribe for term deposits. The recommender system is useful to any business that makes money via recommendations. We have taken the ‘Portuguese banking institution’ dataset for this purpose. Giving good recommendations will help the banks improve their business and also at the same time deliver to the customers as to exactly what they are looking for. This will help the customer to carry on with the best service and have a good experience. We also believe that recommendation systems play a vital role in the banking platforms, where they invest in data science to improve their customer satisfaction by attracting new subscribers to their plans through proper campaigns. This also helps the bank understand the customer churn better.
* Our data includes ,
* Numerical Variables :10
* Categorical Variables : 10

**Dataset and Domain**

**Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **COLUMN NAME** | **COLUMN DESCRIPTION** |
| 1. | age | Age of the user |
| 2. | job | Type of Job |
| 3. | marital | Marital status |
| 4. | education | Educational Background |
| 5. | Default | Credit in default ? |
| 6. | housing | Housing loan? |
| 7. | loan | Personal loan? |
| 8. | contact | Contact communication type |
| 9. | month | Last contact month of the year |
| 10. | day\_of\_week | Last contact day of the week |
| 11. | duration | Last contact duration in seconds |
| 12. | campaign | Number of contacts performed during this campaign |
| 13. | pdays | Number of days that passed by after the client was last contacted from a previous campaign |
| 14. | previous | Number of contacts performed before this campaign and for this client |
| 15. | poutcome | outcome of the previous marketing campaign |
| 16. | emp.var.rate | employment variation rate - quarterly indicator |
| 17. | cons.price.idx | consumer price index - monthly indicator |
| 18. | cons.conf.idx | consumer confidence index - monthly indicator |
| 19. | euribor3m | euribor 3 month rate - daily indicator |
| 20. | nr.employed | number of employees - quarterly indicator |
| 21. | Output Variable : y? | Has the client subscribed to a term deposit? |

**Data Preprocessing**

**Null value treatment**

* df.isnull().sum() :
* job 330
* marital 80
* education 1731
* default 8597
* housing 990
* loan 990

Missing values have been handled. Each null values has been treated based on the percentage of null values when compared to the entire dataset’s value count. Deleted the rows (dropna(axis=0))for value counts that had null values less than 10%. Replaced null values (fillna())with mean/ median for numerical data based on the skewness and symmetry of the data. For categorical data, replaced value based on mode value.

**Feature Engineering**

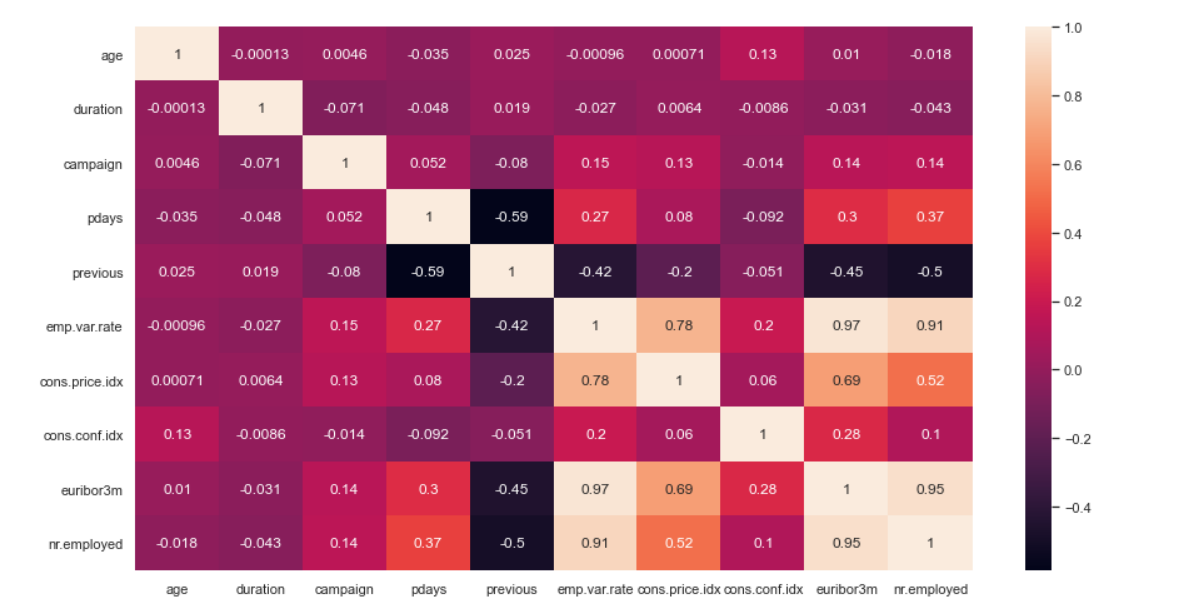
* Standard Scaler method was used to scale the data , this removes the mean and scales each variable to unit variance.
* Scaling and Transformation was performed on numerical data. In the project based on appropriate values, we have used square root , Reciprocal and Boxcox transformation.

**Outlier treatment**

* Based on Box plot , we identified the outliers present and we have treated using Z-score and Inter quartile range. In places where skewness was observed we have used the IQR and to calculate the probability score occurring within normal distribution we have used the Z score method.

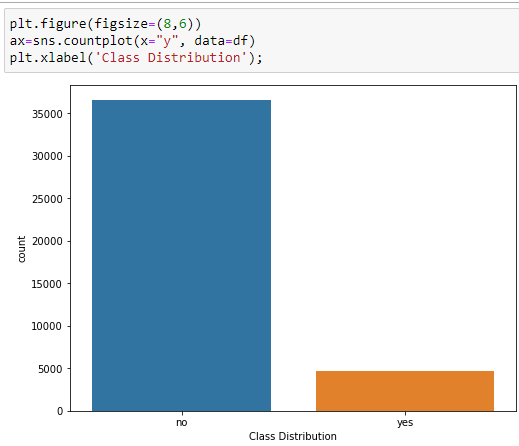
**Multi-collinearity**

* when two or moreindependent variables arecorrelated highly with one another in a regression model multicollinearity occurs.



**Imbalance Data**

* Our target variable, the data is an imbalanced data. As seen below, we observe that the number of subscribers who do not subscribe to term deposit out way that of those who do. Our goal is to predict a better way to reach out to our targeted audience to improve Subscription for deposits.
* Most machine learning algorithms assume data equally distributed. So, when we have a class imbalance, the machine learning classifier tends to be more biased towards the majority class, causing bad classification of the minority class. Here our using confusion matrix it is seen that the dataset is highly unbalanced, with nearly all client actually decline to subscribe.

****

**Salient Features:**

Based on our below stated hypothesis statements ,

How the features are related to target variable ‘y’:

1. How does the age of customer affect subscription?
2. Does having a job influence the decision of subscribing to a bank term deposit ?
3. Does having an existing loan affect the decision of subscribing to a bank term deposit ?
4. Is marital status affecting the customer subscription ?
5. Are social and economic indicators relevant?
6. Are date and time conditions relevant to the subscription ?
7. is term deposit subscription dependent or independent on housing?
8. duration has any effect on term deposit?
9. pdays has any effect on term deposit?
10. nr.employed has any effect on term deposit?
11. contact method has any effect on term deposit?
12. previous has any effect on poutcome?
13. poutcome has effect on subscription

We have identified that from Euribor3m, We can clearly see the difference in median for both the classes. This indicates that the feature can be very useful in predicting the term subscription.

**Step by Step Walk through**

Step 1: Loaded the data frame.

Step 2: Performed Null value Treatment

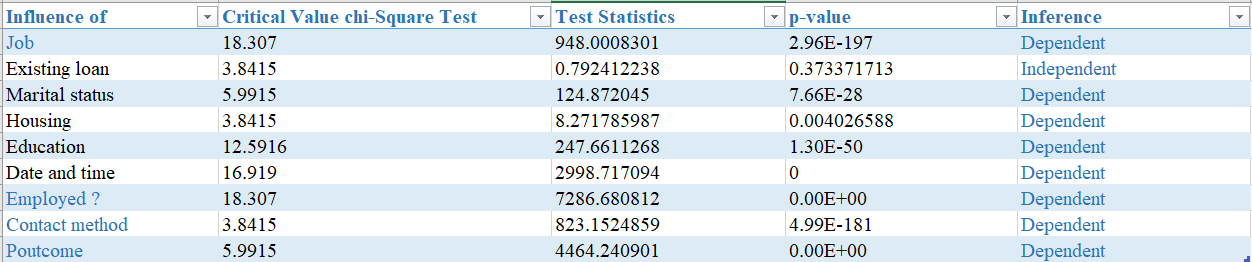
Step 3: Treated categorical data, performed encoding

Step 4: Performed outlier treatment

Step 5: Transformation and scaling

Step 6: Treatment for imbalanced data

Step 7: Identified Significant features



Since most of the features are categorical, used chi2\_contingency test to find the dependency of the independent and dependent variables.

For numeric variables, 2 sample ztest was performed and checked whether mean of the subscription of the variables has any effect on the subscription.

Step 8: Build model

Step 9: Used classification algorithms to further identify the most significant feature that can be used to build a better model