**Statistics 1**

**Hypothesis Testing On Bank Marketing Campaign**

**Manish Kumar**

**D19018**

**Medium link:-** <https://medium.com/@manish.kumar_61520/statistics-1-941486fcdcb9>

**Abstract :-**

**Moro et al., 2014] S. Moro, P. Cortez and P. Rita is a company who gives loan to their customer, and has to return the loan amount as term deposit by per month. Most of the Customer is not giving the term deposit on the given time. Bank has to maximize his profit for that bank want to spread awareness in his customer so bank has organise a marketing campaign. This marketing campaign goes upto Five days, in this period bank employee call to customer and try to convince them to submit the term deposit on time. We have many attributes for our bank customer like their age, educational background , marital status, job, balance etc. We have 45211 customers , we want to conduct a hypothesis test , so we can understand which attributes of customers is important for us so we can maximize our profit.**

**Keywords :- Bank term deposit, Hypothesis testing,**

**Introduction:-**

* **About --** Moro et al., 2014] S. Moro, P. Cortez and P. Rita is a company who wants to make a Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.
* **Marketing campaigns --** are characterized by focusing on the customer needs and their overall satisfaction. Nevertheless, there are different variables that determine whether a marketing campaign will be successful or not. There are certain variables that we need to take into consideration when making a marketing campaign.
* **Problem Statement —**Find the best strategies to improve for the next marketing campaign. How can the financial institution have a greater effectiveness for future marketing campaigns? In order to answer this, we have to analyze the last marketing campaign the bank performed and identify the patterns that will help us find conclusions in order to develop future strategies.
* **Objective –** To understand customer is going to submit the term deposit on time or not for that we have to make a strategies so a customer will submit the term deposit.
* **Dataset description --** Total no. of observation 45211
* **Attribute description :-**
* **Ai. bank client data:**
* 1 - **age:** (numeric)  
  2 - **job:** type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
  3 - **marital:** marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
  4 - **education:** (categorical: primary, secondary, tertiary and unknown)  
  5 - **default:** has credit in default? (categorical: 'no','yes','unknown')  
  6 - **housing:** has housing loan? (categorical: 'no','yes','unknown')  
  7 - **loan:** has personal loan? (categorical: 'no','yes','unknown')

8 - **balance:** Balance of the individual.

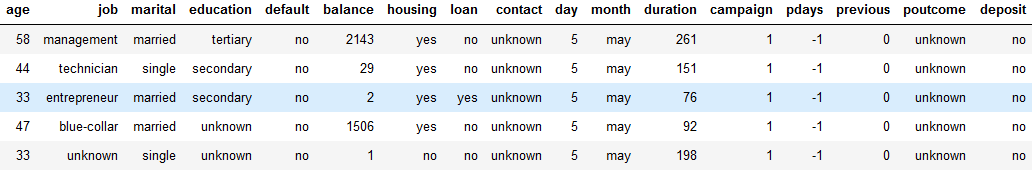
**Related with the last contact of the current campaign**

9 - **contact:** contact communication type (categorical: 'cellular','telephone')   
 10 - **month:** last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
 11 - **day:** last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
 12 - **duration:** last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

**other attributes**

13 - **campaign:** number of contacts performed during this campaign and for this client (numeric, includes last contact)  
 14 - **pdays:** number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
 15 - **previous:** number of contacts performed before this campaign and for this client (numeric)  
 16 - **poutcome:** outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

**Output variable (desired target):** 17 – **deposit** - has the client subscribed a term deposit? (binary: 'yes','no')



**Approach:-**

**In the dataset the target variable deposit is binary. For hypothesis testing separated the independent columns names in** numerical and categorical. For numerical variable used t-test and observed p-values. All numerical variables p-values observed less than 0.05 showing that all of these are important variables with 5% critical values.

Similarly, for categorical variable used chi square test and observed p-values. All categorical variables p-values observed less than 0.05 showing that all of these are important variables with 5% critical values.

**Numerical variables:-**

* **Age :- To check varaible age is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- Age of customer is not dependent on term deposit
* Alternate hypothesis(H1) :- Age of customer is depend on term deposit
* Age of customer is going to impact on marketing campaign means a customer's age is depending on customer's age becuase i got P-value for variable age is 8.825643691922395e-08 that is below than our level of significant(where alpha = 0.05).so here type 1 error occurs so wegoing to reject null hypothesis.
* Type 1 error :- Type 1 error will occur when age do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when age impact marketing campaign but we fail to reject null hypothesis.
* Business Context:- When Age of customer’s is between 23 to 40 they repay the term deposit as well as their age is increasing they are not paying the term deposit some of the customers who are retired they repay the term deposit on time.
* **Balance:- To check variable Balance is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- balance of customer is not dependent on term deposit
* Alternate hypothesis(H1) :- balance of customer is depend on term deposit
* Account Balance of customer is going to impact on marketing campaign means a customer's balance is depending on customer's term deposit because i got P-value for variable age is 8.825643691922395e-08 that is below than our level of significant(where alpha = 0.05).so here type 1 error occurs so we going to reject null hypothesis.
* Type 1 error :- Type 1 error will occur when balance do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when balance impact marketing campaign but we fail to reject null hypothesis
* Business context :- Balance of relavent because we got most of the customer who are not going to repay the loan again their account balance is below than 5000.
* **Day :- To check varaible day is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- day is not dependent on term deposit
* Alternate hypothesis(H1) :- day is depend on term deposit
* day is going to impact on term deposit means a variable day is depending on custemer's term deposit becuase i got P-value for variable age is 1.653880160817521e-09 that is below than our level of significant(where alpha = 0.05).so herr type 1 error occurs so wegoing to reject null hypothesis.
* Type 1 error :- Type 1 error will occur when day do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when day impact marketing campaign but we fail to reject null hypothesis.
* Business context :- This marketing campaign goes till 35 days so we found that most of the people from day 5 to day 23 they not paying the term deposit in the starting of campaign they submitting the term deposit and last of marketing campaign they submitting the term deposit

**Campaign:-To check variable campaign is significant or not for we create a hypothesis**

* Null hypothesis testing(H0) :- campaign is not dependent on term deposit
* Alternate hypothesis(H1) :- campaign is depend on term deposit
* campaign is going to impact on term deposit means a variable campaign is depending on customer's term deposit becuase i got P-value 1.0123473642096297e-54 for variable campaign that is below than our level of significant(where alpha = 0.05).so here type 1 error occurs so we going to reject null hypothesis.
* Type 1 error :- Type 1 error will occur when campaign do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when campaign impact marketing campaign but we fail to reject null hypothesis.
* Business context:-campaign is going for 5 days that shows that which days how many customers are attracted with bank employee

**Pdays :-To check varaible padys is significant or not for we create a hypothesis**

* Null hypothesis testing(H0) :- Pdays is not dependent on term deposit
* Alternate hypothesis(H1) :- Pdays is depend on term deposit
* pdays is going to impact on term deposit means a variable pdays is depending on customer's term deposit becuase i got P-value 3.790553319478034e-108 for variable pdays that is below than our level of significant(where alpha = 0.05).so here type 1 error occurs so we going to reject null hypothesis.
* Type 1 error :- Type 1 error will occur when pdays do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when pdays impact marketing campaign but we fail to reject null hypothesis
* Business context:- Pdyas is a campaigning attribute which shows that bank employee has contacted how many customers.
* **Previous :-To check varaible previous is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- Previous is not dependent on term deposit
* Alternate hypothesis(H1) :- Previous is depend on term deposit
* previous is going to impact on term deposit means a variable previous is depending on custemer's term deposit becuase i got P-value 7.801830349805211e-88 for variable previous that is below than our level of significant(where alpha = 0.05).so here type 1 error occurs so we going to reject null hypothesis
* Type 1 error :- Type 1 error will occur when previous do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when previous impact marketing campaign but we fail to reject null hypothesis..
* Business context:- it shows how many customers are connected previously before campaign

**Categorical variable**

**Job :- To check varaible job is significant or not for we create a hypothesis**

* Null hypothesis testing(H0) :- job is not dependent on term deposis
* Alternate hypothesis(H1) :- job is dependent on term deposit
* In this hypothesis test we want to check the given variable job is going to impact on term deposit,after doing chi-square test we found that our p-value is 0.0000 and carmer's v is 0.1360 so we going to reject the null hypothesis and here is type 1 error occurs.
* Cramer's V is a way of calculating correlation in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance.
* Type 1 error :- Type 1 error will occur when job do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when job impact marketing campaign but we fail to reject null hypothesis**.**
* **Business context:-**

**Marital :-To check varaible previous is significant or not for we create a hypothesis**

* Null hypothesis testing(H0) :- marital is not dependent on term deposit
* Alternate hypothesis(H1) :- marital is dependent on term deposit
* In this hypothesis test we want to check the given variable marital is going to impact on term deposit,after doing chi-square test we found that our p-value is 0.0000 and carmer's v is 0.1360 so we going to reject the null hypothesis and here is type 1 error occurs in this scenario.
* Cramer's V is a way of calculating correlation in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance.
* Type 1 error :- Type 1 error will occur when martial do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when martial impact marketing campaign but we fail to reject null hypothesis.
* **Education:-To check varaible education is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- education is not dependent on term deposit
* Alternate hypothesis(H1) :- education is dependent on term deposit
* In this hypothesis test we want to check the given variable education is going to impact on term deposit,after doing chi-square test we found that our p-value is 0.0000 and carmer's v is 0.0727 so we going to reject the null hypothesis and here is type 1 error occurs in this scenario.
* Cramer's V is a way of calculating correlation in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance.
* Type 1 error :- Type 1 error will occur when martial do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when martial impact marketing campaign but we fail to reject null hypothesis.
* **Housing :-To check varaible previous is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- housing is not dependent on term deposit
* Alternate hypothesis(H1) :- housing is dependent on term deposit
* In this hypothesis test we want to check the given variable housing is going to impact on term deposit,after doing chi-square test we found that our p-value is 0.0000 and carmer's v is 0.1392 so we going to reject the null hypothesis and here is type 1 error occurs in this scenario.
* Cramer's V is a way of calculating correlation in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance.
* Type 1 error :- Type 1 error will occur when education do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when education impact marketing campaign but we fail to reject null hypothesis.

**Loan :-To check varaible loan is significant or not for we create a hypothesis**

* Null hypothesis testing(H0) :- variable loan is not dependent on term deposit
* Alternate hypothesis(H1) :- variable loan is dependent on term deposit
* In this hypothesis test we want to check the given variable loan is going to impact on term deposit,after doing chi-square test we found that our p-value is 0.0000 and carmer's v is 0.0682 so we going to reject the null hypothesis and here is type 1 error occurs in this scenario.
* Cramer's V is a way of calculating correlation in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance.
* **duration :- To check varaible duration is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- call duration is not impacting on marketing campaign
* Alternate hypothesis(H1) :- call duration is impacting on marketing campaign
* call duration is giving impact on marketing campaign means a customer is attracted during marketing campaign so he is going to give deposit and
* Type 1 error :- Type 1 error will occur when call duration do not impact marketing campaign but we falsely reject null hypothesis
* Type 2 error :- Type 2 error will occur when call duration impact marketing campaign but we fail to reject null hypothesis.
* Business Context :- Call duration is used to influence the customer so will submit the term deposit here most of the customer is not influence by bank employee so they are not paying the term deposit on time.
* **Contact:-To check varaible previous is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- varaiable contact is not dependent on term deposit
* Alternate hypothesis(H1) :- variable contact is dependent on term deposit
* In this hypothesis test we want to check the given variable contact is going to impact on term deposit,after doing chi-square test we found that our p-value is 0.0000 and carmer's v is 0.1514 so we going to reject the null hypothesis and here is type 1 error occurs in this scenario.
* Cramer's V is a way of calculating correlation in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance.
* **Month:-To check varaible previous is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- variable month is not dependent on term deposit
* Alternate hypothesis(H1) :- variable month is dependent on term deposit
* in this hypothesis test we want to check the given variable month is going to impact on term deposit, after doing chi-square test we found that our p-value is 0.0000 and carmer's v is 0.2602 so we going to reject the null hypothesis and here is type 1 error occurs in this scenario.
* Cramer's V is a way of calculating correlation in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance.
* **Potcome:-To check varaible previous is significant or not for we create a hypothesis**
* Null hypothesis testing(H0) :- poutcome is not dependent on term deposit
* Alternate hypothesis(H1) :- poutcome is dependent on term deposit
* In this hypothesis test we want to check the given variable education is going to impact on term deposit,after doing chi-square test we found that our p-value is 0.0000 and carmer's v is 0.3117 so we going to reject the null hypothesis and here is type 1 error occurs in this scenario.
* Cramer's V is a way of calculating correlation in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance.

**Conclusion:-**

**Two sample T- test**

|  |  |  |  |
| --- | --- | --- | --- |
| **index** | **Variable name** | **T-statistics** | P-value |
| 1 | duration | 57.5141 | 0 |
| 2 | campaign | -22.8007 | 3.73E-112 |
| 3 | pdays | 18.9435 | 7.33E-78 |
| 4 | previous | 18.118 | 1.36E-71 |
| 5 | balance | 9.93355 | 4.38E-23 |
| 6 | day | -5.91698 | 3.44E-09 |
| 7 | age | 4.31832 | 1.60E-05 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Categorical variable:-**

**Chi-square Test**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **index** | **Variable name** | **chi-square** | **P-value** | **Carmer's-v** |
| 1 | job | 836.106 | 3.34E-172 | 0.136 |
| 2 | martial | 196.496 | 2.15E-43 | 0.0659 |
| 3 | education | 238.924 | 1.63E-51 | 0.0727 |
| 4 | housing | 875.693 | 2.92E-192 | 0.1392 |
| 5 | loan | 210.195 | 1.67E-47 | 0.0682 |
| 6 | contact | 1035.71 | 1.25E-225 | 0.1514 |
| 7 | month | 3061.84 | 0.00E+00 | 0.26 |
| 8 | poutcome | 4391.51 | 0.00E+00 | 0.3117 |
| 9 | default | 22.7235 | 2.45E-06 | 0.0224 |
|  |  |  |  |  |
|  |  |  |  |  |

**Here we got our all variable is important because our all of variable is below than our level of significance (alpha).**

**Findings :- in this hypothesis testing I got that bank has need all the attributes that is given in the dataset for influence more customer for their term deposit.**

**References:-** <https://www.pythonfordatascience.org/chi-square-test-of-independence-python/>