# Topic: Naive Bayes

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

Name: **Nithin Dsouza**

Batch Id: **:**  050121-10AM

Topic: Naïve Bayes

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**2.1 Data Cleaning, Feature Engineering, etc.**

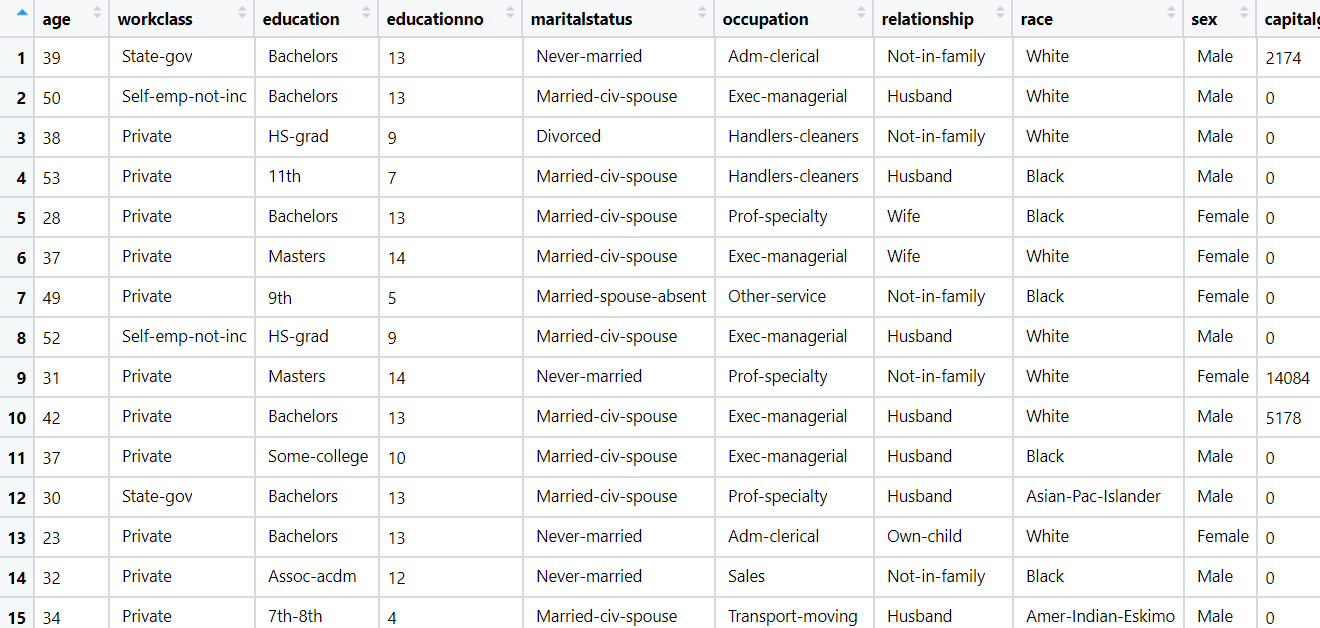
1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform the Naïve Bayes Machine Learning Algorithm**
   3. **Validate the train and test data and perform confusion matrix, get precision, recall and accuracy from it.**
3. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

**Note:**

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the modules (elaborating on steps mentioned above)

1.) Prepare a classification model using Naive Bayes for Salary dataset, train and test datasets are given separately use both datasets for model building.



Pro

**Business Problem**

**Objective :- predict salary**

1. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| Age | **Age of person** | **Ordinal** | **It is relative to experience, so use full** |
| Work class | **Salary depends on class of work** | **Ordinal** | **Use full information, because different organizations have different pay scales** |
| Education | **Education of person** | **Ordinal** | **Relevant information, If well educated, more valuable work, then more salary** |
| Education no | **Level of education** | **Ordinal** | **Use full information, different level of education is numbered different and paid different** |
| Marital status | **Person married status** | **Nominal** | **Not use full** |
| Occupation | **Job of person** | **Ordinal** | **Use full, different job paid different amount** |
| Relationship | **Relationship with parents** | **Ordinal** | **Not use full, salary is independent of relation** |
| Race | **Locality of person** | **Nominal** | **Not use full** |
| Sex | **Gender of person** | **Nominal** | **Use full, because Capacity of doing work, and type of work** |
| Capital gain | **Capital gain of person** | **Ratio** | **Not use full** |
| Capital loss | **Capital loss of person** | **Ratio** | **Not use full** |
| Hours per week | **Working hours per week** | **Ratio** | **Use full, if working hour is more, pay will more** |
| Native | **Nativity of person** | **Nominal** | **Not use full** |
| Salary | **Salary of person** | **Ratio** | **Use full, it is the output variable** |
|  |  |  |  |

1. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| **User ID** | **User ID of Social media user** | **Nominal** | **Not Use full** |
| Gender | **Mae/female** | **Nominal** | **Use full** |
| Age | **Age of social media account holder** | **Ordinal** | **Use full, depends on the age we can predict, 15 age guy has very less chance to buy car** |
| Estimated Salary | **Salary of person** | **Ratio** | **Use full, more salary than more chance of buying** |
| Purchased | **Prediction column** | **Binary** | **Output variable, it depends on all the above inputs** |

**Python code details :**

Here we have two data frame one is train\_df(30161\*14) another is test\_df(15060\*14) as , train and test datasets are given separately use both datasets for model building so we create a new data frame with add both df in one Data Frame name as df. It has 45221 rows and 14 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df and convert ' >50K' to 0 and <=50K to 1 for salary columns. Check for data types ,unique value and variance .

Then we check for unique value in each columns

:-

age 74

workclass 7

education 16

educationno 16

maritalstatus 7

occupation 14

relationship 6

race 5

sex 2

capitalgain 121

capitalloss 97

hoursperweek 96

native 40

Salary 2

Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .

Column “capitalgain","capitalloss","hoursperweek “ has maximum 0 is available so we build our model without those columns .

We have created dataframe with only with (discrete,continuous ,output) and Normalized data frame (considering the numerical part of data) , dummies for discrete data, creating a new df with both df named as X.

**5. Model Building**

**5.1 Build the model on the scaled data (try multiple options)**

**5.2 Perform the Naïve Bayes Machine Learning Algorithm**

**5.3 Validate the train and test data and perform confusion matrix, get precision, recall and accuracy from it.**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 75% data on train and 25% test . Preparing a naive bayes model on training data set ,then test on test data , Evaluation on Test Data as result 78% are right prediction .

Crosstab

Salary <=50K >50K

row\_0

<=50K 6846 711

>50K 1679 2070

Evaluation on Train Data also as result 78% are right prediction .

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Problem Statement: -

This dataset contains information of users in social network. This social network has several business clients which can put their ads on social network and one of the Client has a car company who has just launched a luxury SUV for ridiculous price. Build the Bernoulli Naïve Bayes model using this dataset and classify which of the users of the social network are going to purchase this luxury SUV.

Purchased: - 1 and Not Purchased: - 0

A screenshot of a cell phone

Description automatically generated

**Business Problem**

**Objective :-** Purchased or not

1. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| **ID** | **ID of tweeter account holder** | **Nominal** | **Not use full for analysis** |
| **Keyword** | **Key word tweet contains** | **Categorical** | **Use full for analysis** |
| **Location** | **Location of Tweeter** | **Nominal** | **Use full for analysis** |
| **Text** | **Text of tweets** | **Categorical** | **Use full for analysis** |
| **Target** | **Output column** | **Binary** | **Output variable, depends on input variable, it changes** |

**Python code details :**

Data Frame name is df. It has 400 rows and 5 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df .Check for data types ,unique value and variance .

Then we check for unique value in each columns

:-

User ID 400

Gender 2

Age 43

EstimatedSalary 117

Purchased 2

Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .

We have created dataframe with only with (discrete,continuous ,output) and Normalized data frame (considering the numerical part of data) , dummies for discrete data, creating a new df with both df named as X.

**5. Model Building**

**5.1 Build the model on the scaled data (try multiple options)**

**5.2 Perform the Naïve Bayes Machine Learning Algorithm**

**5.3 Validate the train and test data and perform confusion matrix, get precision, recall and accuracy from it.**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 75% data on train and 25% test . Preparing a naive bayes model on training data set ,then test on test data , Evaluation on Test Data as result 65% are right prediction .

Crosstab

Purchased 0 1

row\_0

0 65 35

Evaluation on Train Data also as result 64% are right prediction .

Problem Statement: -

In this case study you have been given with tweeter data collected from an anonymous twitter handle, with the help of Naïve Bayes algorithm predict a given tweet is Fake or Real about real disaster occurring.

Real tweet: - 1 and Fake tweet: - 0

A screenshot of a cell phone

Description automatically generated

**Business Problem**

**Objective :-** Purchased or not

**Python code details :**

Data Frame name is tweet. It has 7613 rows and 5 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Here we have done text cleaning for text columns .as predictor we have target columns , given tweet is Fake or Real about real disaster occurring

After cleaning the text data we have created corpus and then TDM.

**5. Model Building**

**5.1 Build the model on the scaled data (try multiple options)**

**5.2 Perform the Naïve Bayes Machine Learning Algorithm**

**5.3 Validate the train and test data and perform confusion matrix, get precision, recall and accuracy from it.**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 75% data on train and 25% test . Preparing a naive bayes model on training data set ,then test on test data , Evaluation on Test Data as result 80% are right prediction .

Crosstab

target 0 1

row\_0

0 985 283

1 96 540

Evaluation on Train Dataalso as result 90% are right prediction .

Then we have tried model with different alpaha

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_END\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_