**INTERVIEW PREDICTION**

**DOCUMENTATION**

**PREDICTION OF INTERVIEW**

**ATTENDANCE USING**

**MACHINE LEARNING AND PYTHON**

**ANIKET CHATTOPADHYA, RCCIIT, REG NO. 161170110007**

**CAMELIA MAHATO, RCCIIT, REG NO.161170110025**

**CHANDRABALI BISHNU, RCCIIT, REG NO.:161170110026**

**PALLAB CHAKRABORTY, RCCIIT, REG NO.:161170110048**

**SOHAM MANDAL, RCCIIT, REG NO.161170110070**

**TABLE OF CONTENTS**

1. ACKNOWLEDGEMENT
2. PROJECT OBJECTIVE
3. PROJECT SCOPE
4. DATA DESCRIPTION
5. DATA LOADING
6. DISTRIBUTION ANALYSIS
7. USER DEFINED FUNCTIONS
8. DATA CLEANING
9. ONE HOT ENCODING
10. MODEL BUILDING
11. CONCLUSION
12. CODE

**ACKNOWLEDGEMENT**

I take this opportunity to express my profound gratitude and deep regards to my faculty Mr. Titas Roy Chowdhury for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessing, help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

I am obliged to my project team members for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my project.

**PROJECT OBJECTIVE**

Our project objective was to analyze a data sheet on Interview prediction stats consisting around twenty eight data entries. We analyzed the data using various parameters and tried to cover almost every aspect of the given data and using Machine Learning and python programming. We represented our analysis in both statistically and pictorially. Finally we applied various models on the data, selected the best feature and calculated precision of the data.

**DATA DESCRIPTION**

**Data Info:**

df.info()

df.describe()

df.index()

df.columns()

Range Index : 1234 entries

Data columns : 28 columns

**Column** **Renamed**

Date of interview doi

Client name cl\_nam

Industry indus

Location cl\_loc

Position to be closed pos

Nature of Skillset skill

Interview Type intrvw\_typ

Name(Cand ID) cand\_nam

Gender gend

Candidate Current Location cand\_cur\_loc

Candidate Job Location cand\_j\_loc

Interview Venue intrvw\_ven

Candidate Native Location cand\_nat\_loc

Have you obtained the enq\_perm

necessary permission to start at

the required time

Hope there will be no unscheduled enq\_unsch\_meet

meetings

Can I call you three hours before enq\_call

the interview and follow up on

your attendance for the interview

Can I have an alternative enq\_num

number/desk number. I assure

you that I will not trouble you

too much

Have you taken a printout of enq\_resume

your updated resume. Have you

read the JD and understood the

same

Are you clear with the venue enq\_ven

details and the landmark

Has the call letter been enq\_call\_letter

shared

Expected Attendance expc\_at

Observed Attendance obs\_at

Marital Status married

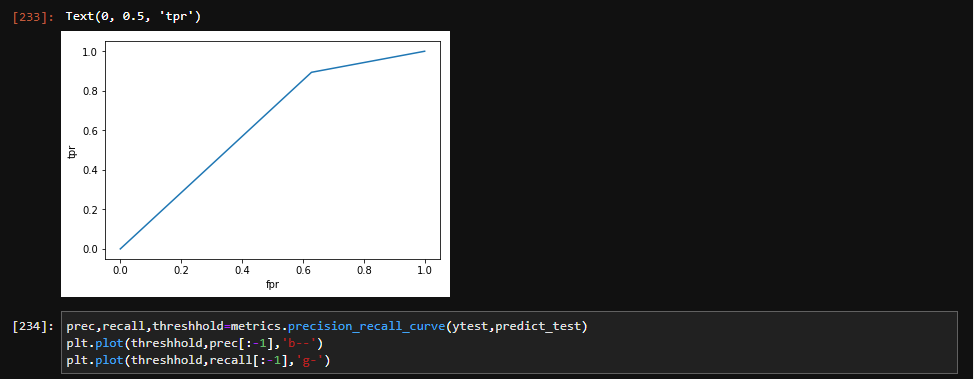
**DATA LOADING**

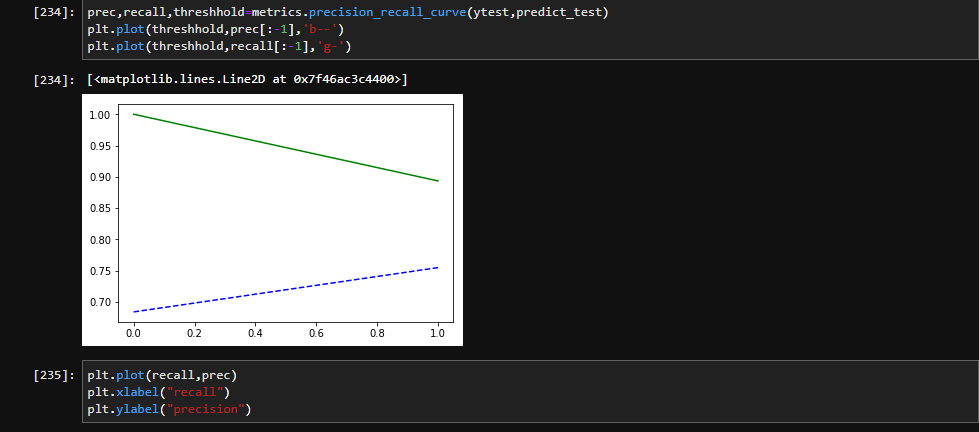
In order to access the data file in the python interpreter we first created a data frame in python using the provided data sheet and loaded the data in the interpreter using the data frame.

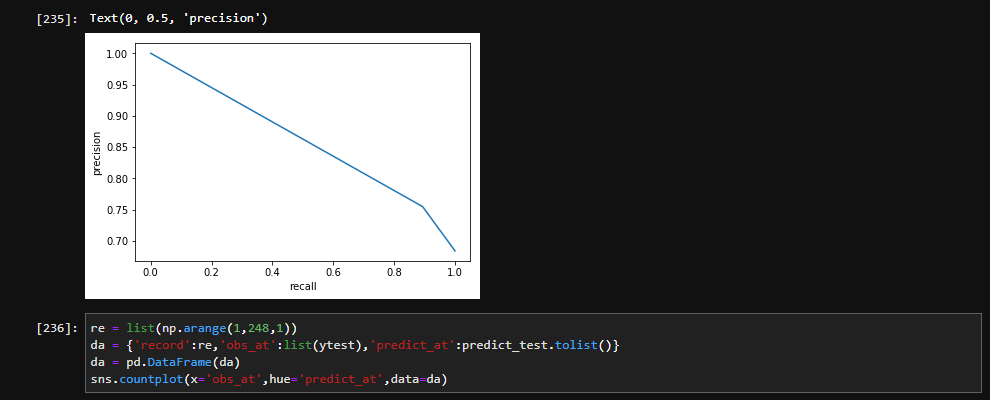
The python code for this process is given below:

**df=pd.read\_csv( “Interview.csv”)**

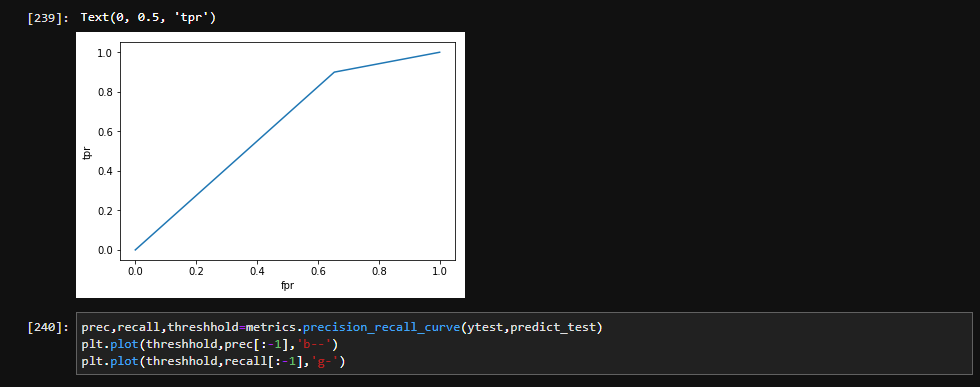
**DISTRIBUTION ANALYSIS**

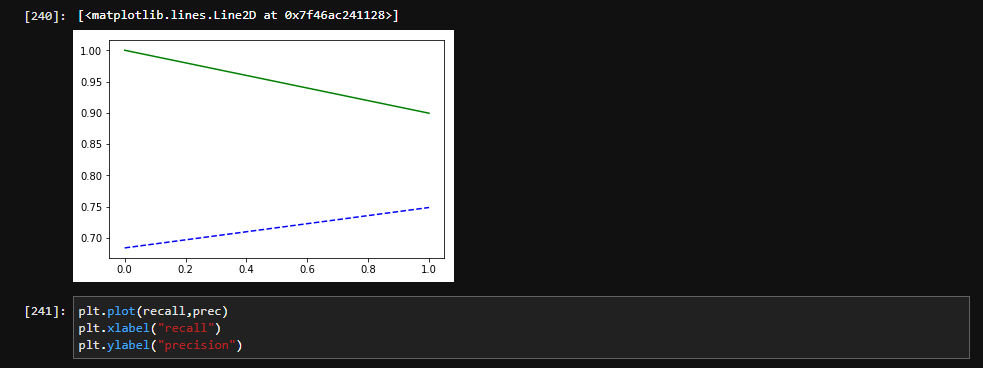


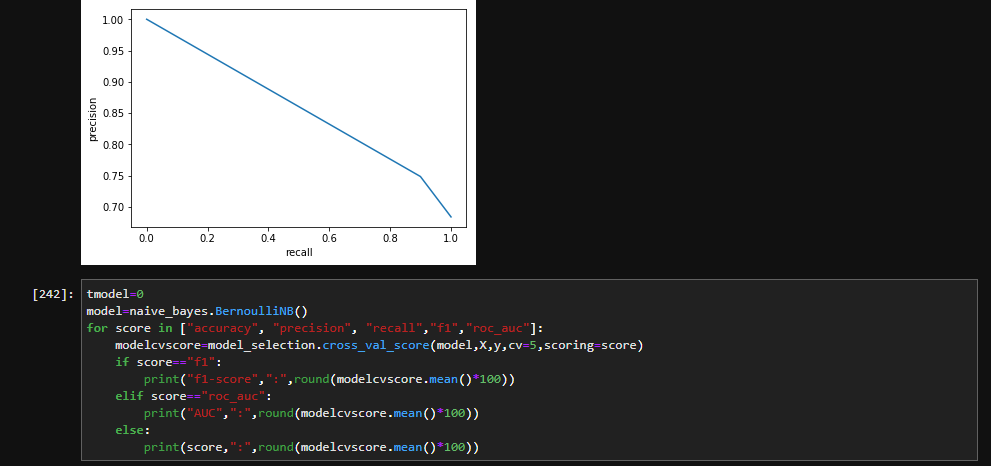


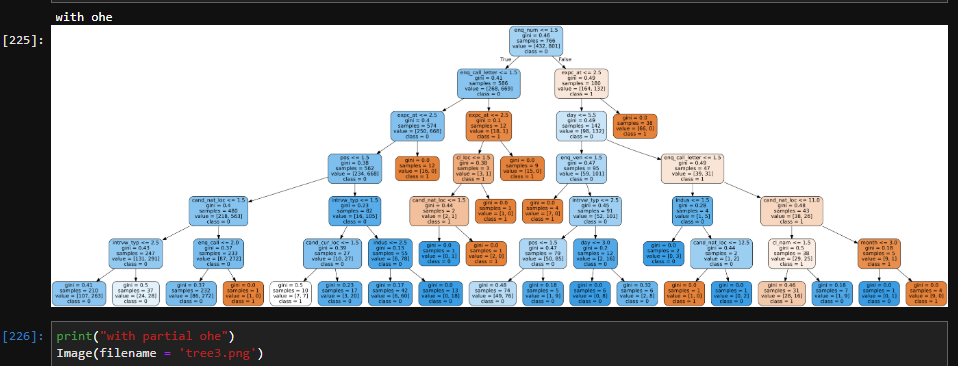


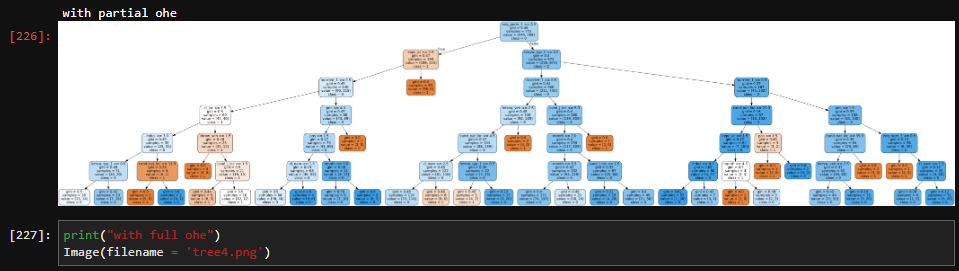


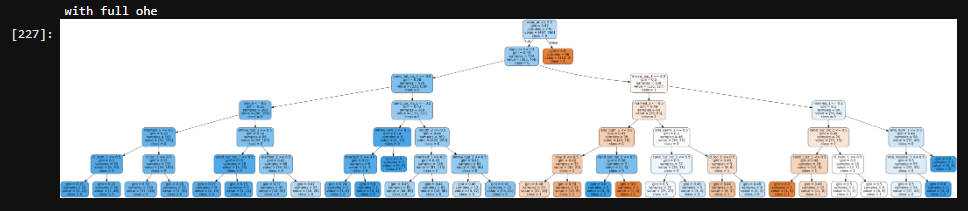


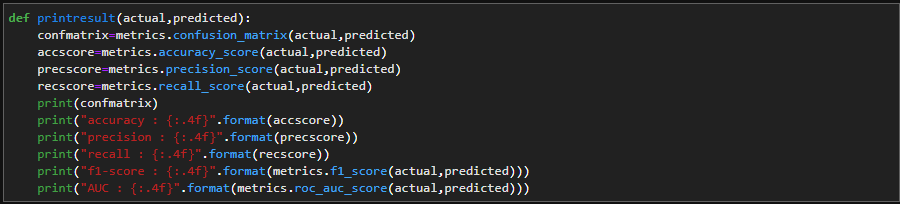










**User Defined Functions Used:**

Here a function is defined named “printresult”. In it we pass actual and predicted result and gives as output confusion matrix and scores of accuracy, precision,recall,f1-score and auc.

DATA CLEANING

**Data cleaning** or **data cleansing** is the process of detecting and correcting (or removing) corrupt or inaccurate [records](https://en.wikipedia.org/wiki/Storage_record) from a record set, [table](https://en.wikipedia.org/wiki/Table_(database)), or [database](https://en.wikipedia.org/wiki/Database) and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the [dirty](https://en.wikipedia.org/wiki/Dirty_data) or coarse data.[[1]](https://en.wikipedia.org/wiki/Data_cleansing#cite_note-1) Data cleansing may be performed [interactively](https://en.wikipedia.org/wiki/Interactively) with [data wrangling](https://en.wikipedia.org/wiki/Data_wrangling) tools, or as [batch processing](https://en.wikipedia.org/wiki/Batch_processing) through [scripting](https://en.wikipedia.org/wiki/Script_(computing)).

After cleansing, a [data set](https://en.wikipedia.org/wiki/Data_set) should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different [data dictionary](https://en.wikipedia.org/wiki/Data_dictionary) definitions of similar entities in different stores. Data cleaning differs from [data validation](https://en.wikipedia.org/wiki/Data_validation) in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

The actual process of data cleansing may involve removing [typographical errors](https://en.wikipedia.org/wiki/Typographical_error) or validating and correcting values against a known list of entities. The validation may be strict (such as rejecting any address that does not have a valid [postal code](https://en.wikipedia.org/wiki/Postal_code)) or [fuzzy](https://en.wikipedia.org/wiki/Fuzzy_logic) (such as correcting records that partially match existing, known records). Some data cleansing solutions will clean data by cross checking with a validated data set. A common data cleansing practice is data enhancement, where data is made more complete by adding related information. For example, appending addresses with any phone numbers related to that address. Data cleansing may also involve activities like, harmonization of data, and standardization of data. For example, harmonization of short codes (st, rd, etc.) to actual words (street, road, etcetera). Standardization of data is a means of changing a reference data set to a new standard, ex, use of standard codes.

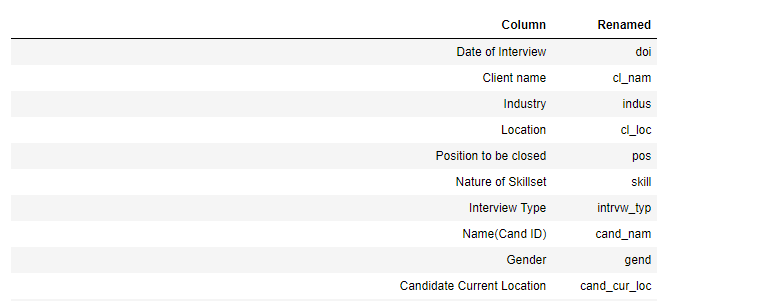
Data cleaning can also be defined as the process of altering data in a given storage resource to make sure that it is accurate and correct. There are many ways to pursue data cleaning in various software and data storage architectures; most of them center on the careful review of data sets and the protocols associated with any particular data storage technology.

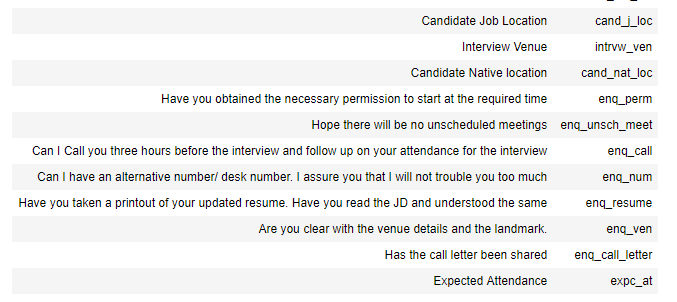
Data cleansing is sometimes compared to data purging, where old or useless data will be deleted from a data set. Although data cleansing can involve deleting old, incomplete or duplicated data, data cleansing is different from data purging in that data purging usually focuses on clearing space for new data, whereas data cleansing focuses on maximizing the accuracy of data in a system. A data cleansing method may use parsing or other methods to get rid of syntax errors, typographical errors or fragments of records. Careful analysis of a data set can show how merging multiple sets led to duplication, in which case data cleansing may be used to fix the problem.

Many issues involving data cleansing are similar to problems that archivists, database admin staff and others face around processes like data maintenance, targeted data mining and the extract, transform, load (ETL) methodology, where old data is reloaded into a new data set. These issues often regard the syntax and specific use of command to effect related tasks in database and server technologies like SQL or Oracle. Database administration is a highly important role in many businesses and organizations that rely on large data sets and accurate records for commerce or any other initiative.

The data cleaning was done by the below mentioned processes:-

1. Renaming column names







The data sheet have 1234 entries, where 28 columns are present. We renamed the column names to make it look more accurate.

For example,

**Column** **Renamed**

Date of interview doi

Client name cl\_nam

Industry indus

Location cl\_loc

Can I call you three hours before enq\_call

the interview and follow up on

your attendance for the interview

The Date of interview is renamed to “doi”, Client name to “cl\_nam”, Industry to “indus”, Location to “cl\_loc” i.e, client location . The columns having question names like Can I call you three hours before were renamed as an enquiry call as “enq\_call” and so on.

1. Removing spaces

Removing of unnecessary spaces have been done while the cleansing of data.

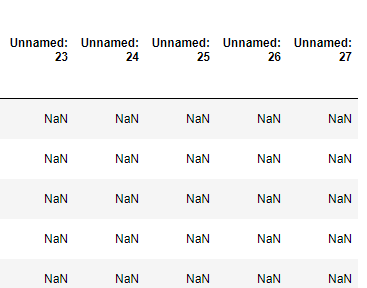
This was done using the str.split() function which is an inbuilt function in python.

1. Unnamed columns dropped

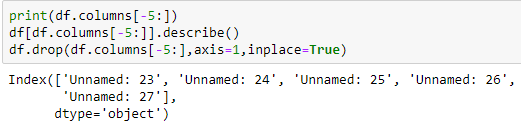
The columns that are unnamed in the data sheet are dropped. There are total five unnamed columns.

Unnamed 23, unnamed 24, unnamed 25, unnamed 26 and unnamed 27.

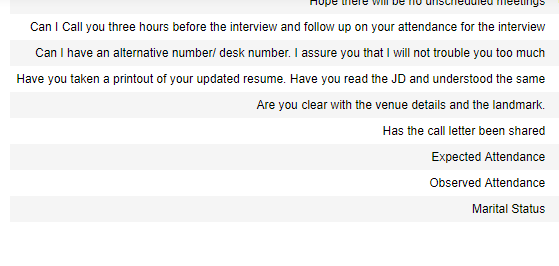
Before:



Code:



Output:

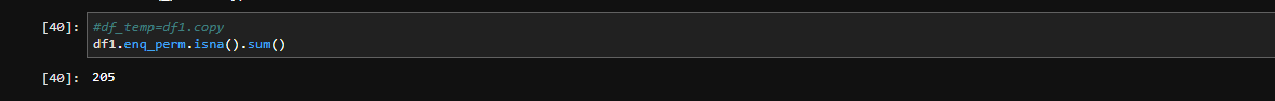


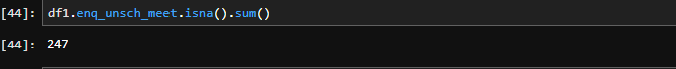
1. Handling of null value

All the columns were checked and we have searched for null values if any. We found the total number of null datas. Then we handled those datas by replacing the null values with some other values mostly with a “no” in our project. In the end there were no null values and the dataset was ready to be modeled.

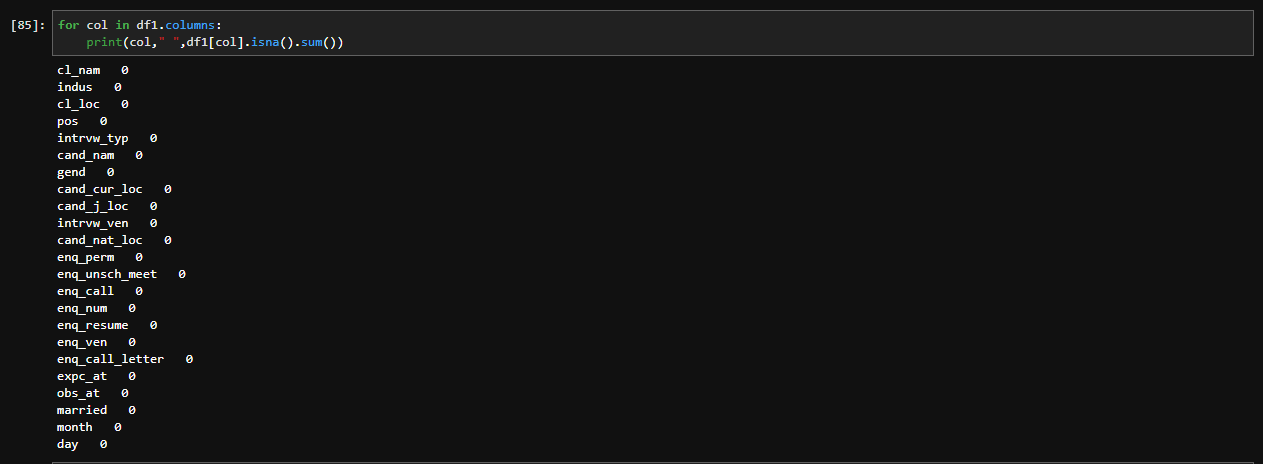
Before:

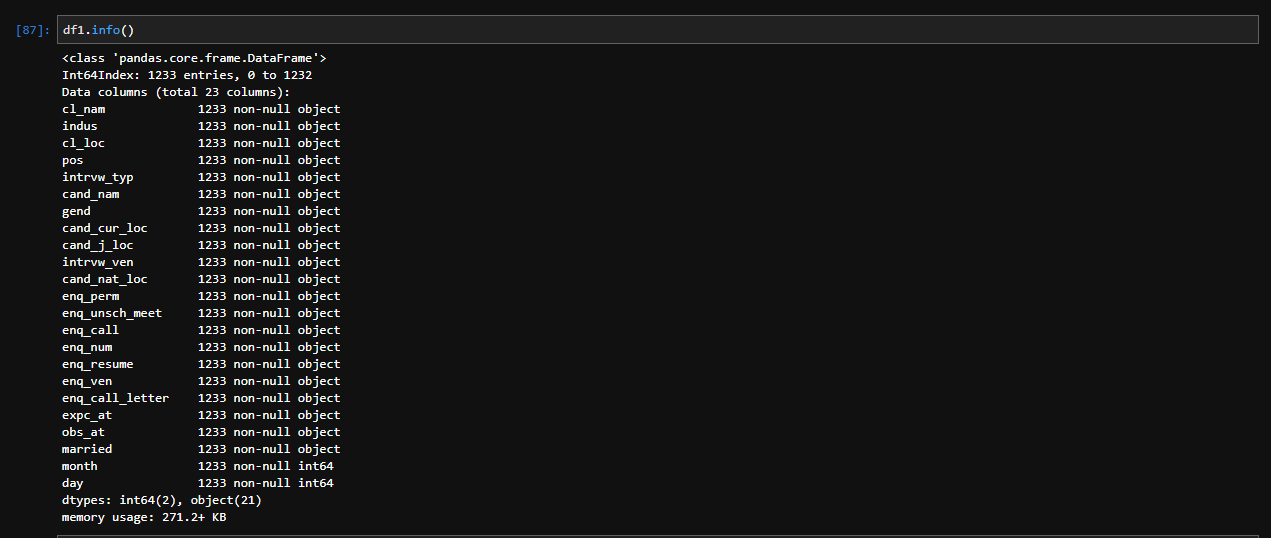
For some columns total sum value was calculated-





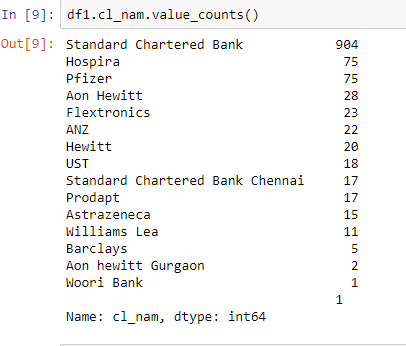
After:



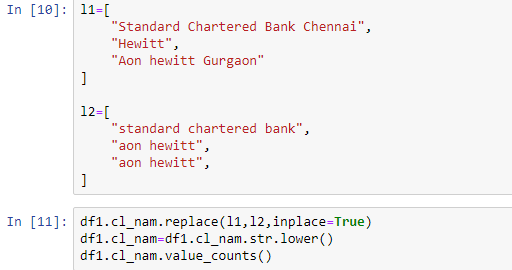


1. Categorising

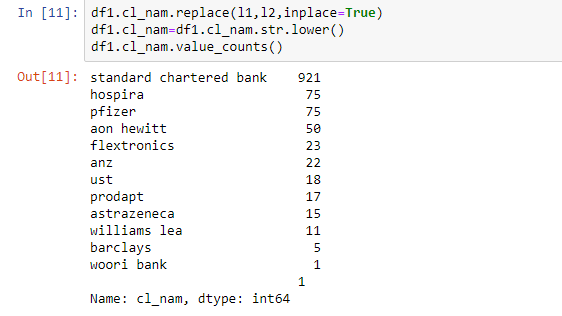
Before:



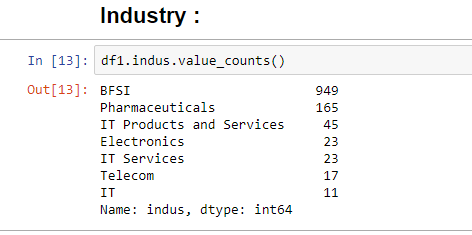
After:



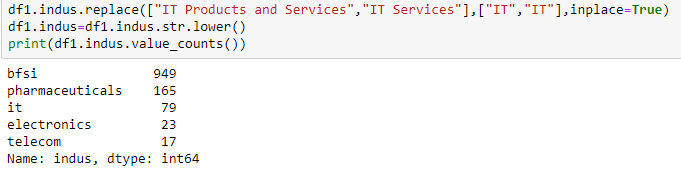
Output:



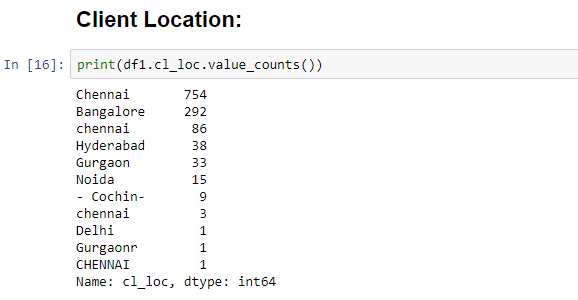
Before:



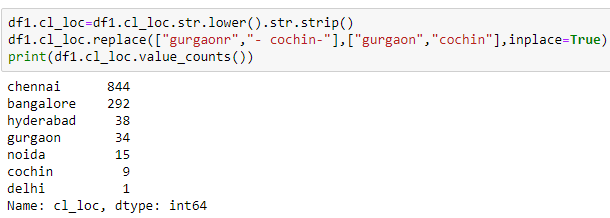
After:



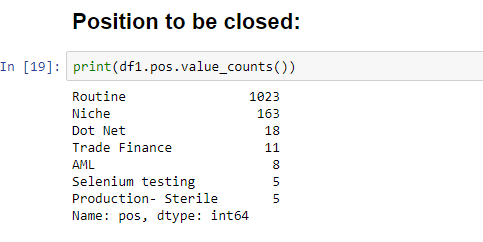
Before:



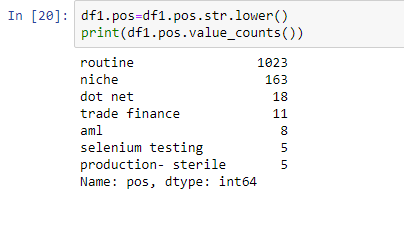
After:



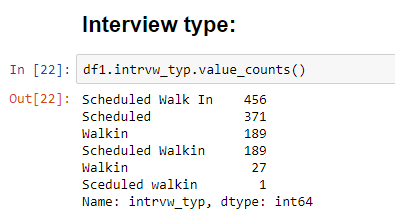
Before:



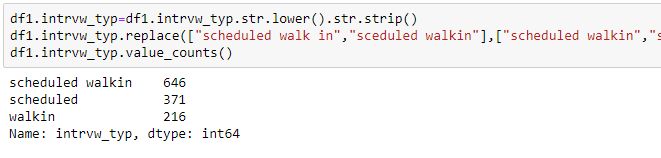
After:



Before:

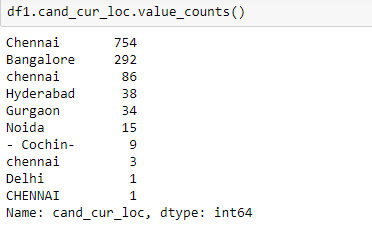


After:

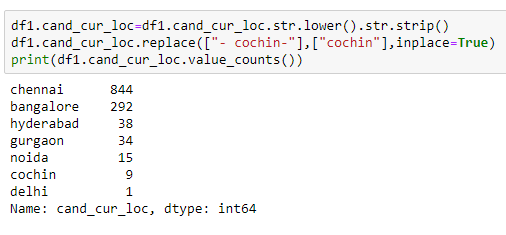


Before:



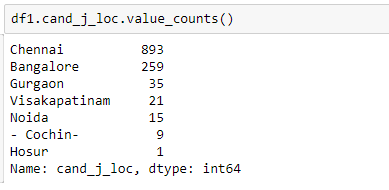


After:

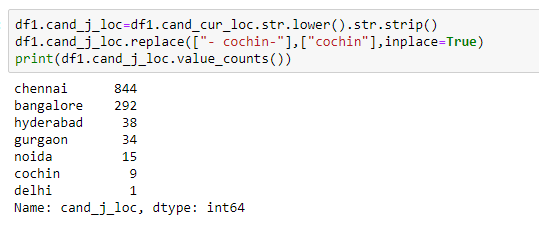




Before:

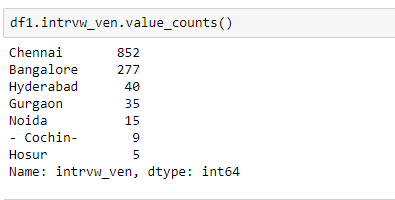


After:

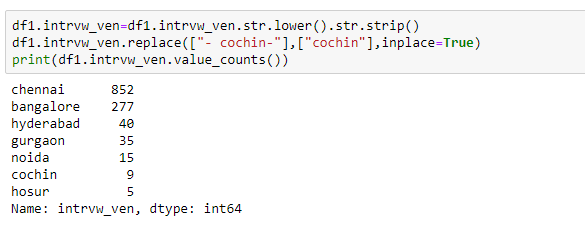




Before:

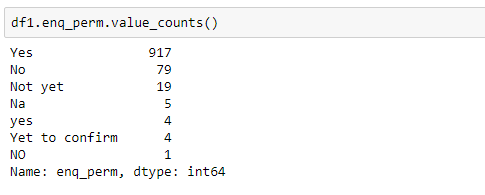


After:

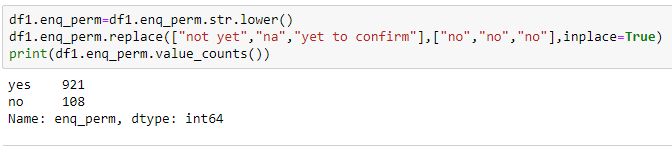




Before:

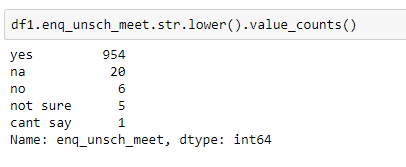


After:

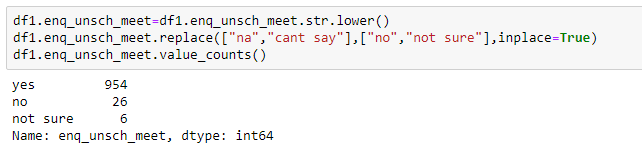




Before:

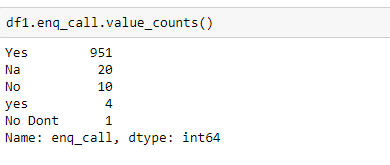


After:

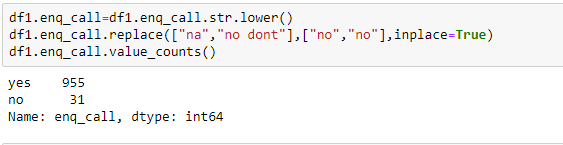




Before:

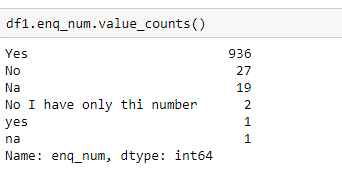


After:

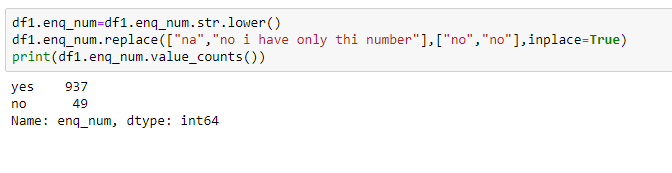




Before:

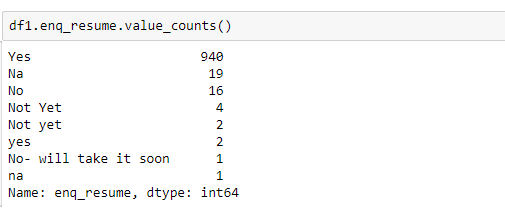


After:

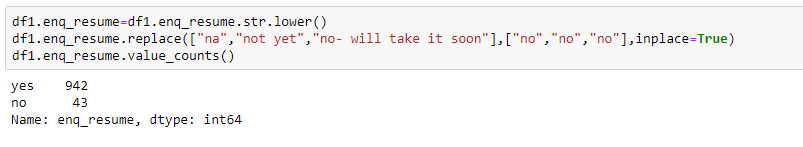




Before:

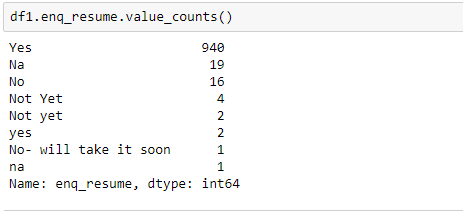


After:

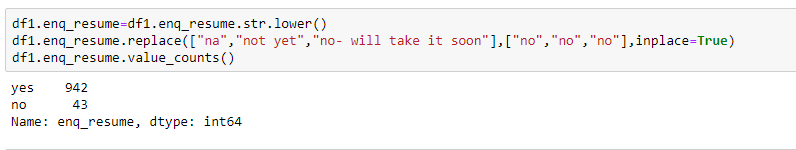




Before:

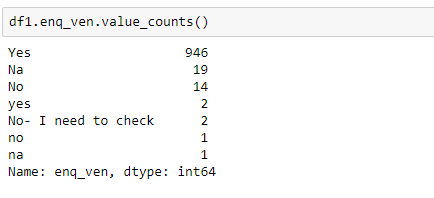


After:

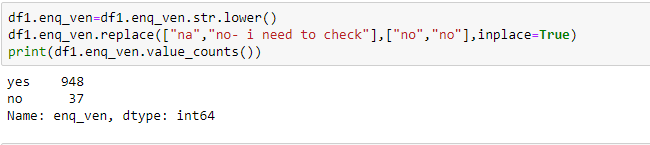




Before:

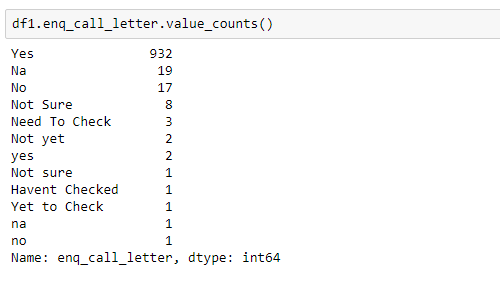


After:

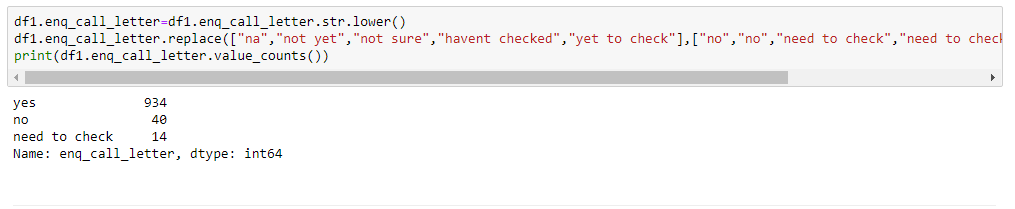


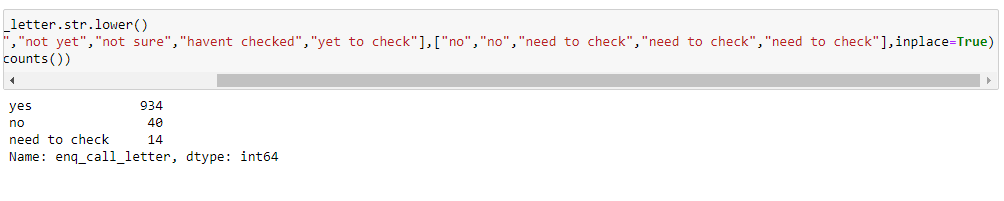


Before:



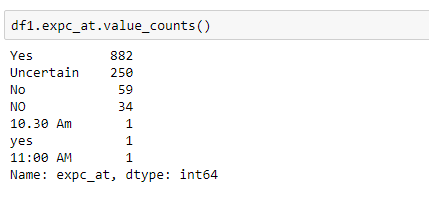
After:



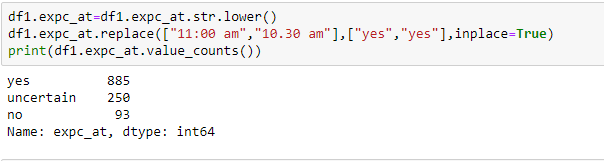




Before:

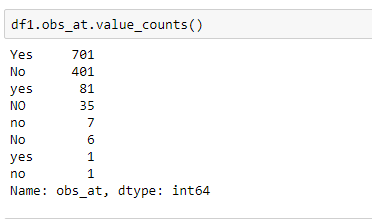


After:

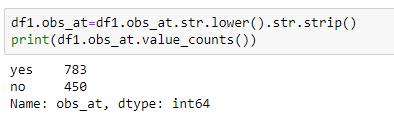




Before:

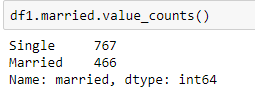


After:

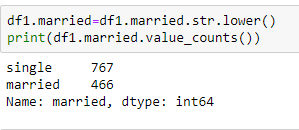




Before:

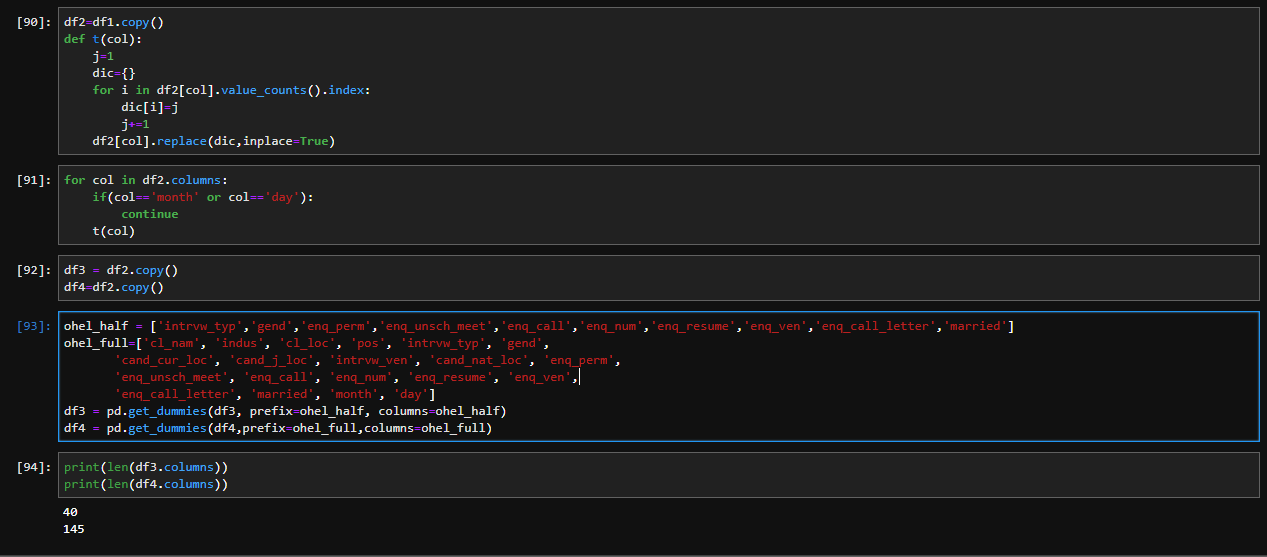


After:



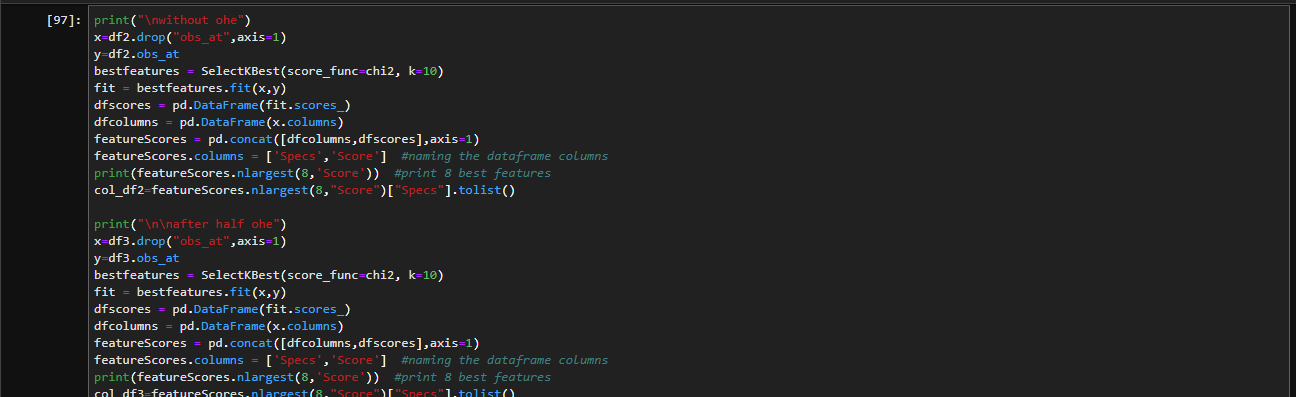
**ONE HOT ENCODING**

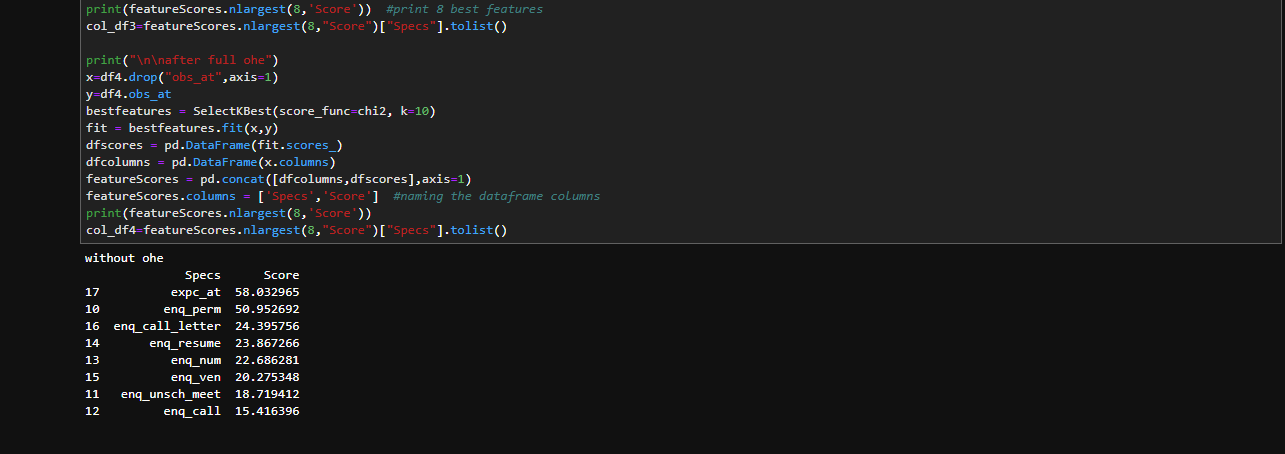
One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. Our categories were formerly rows, but now they’re columns. Our numerical variable, calories, has however stayed the same. A 1 in a particular column will tell the computer the correct category for that row’s data. In other words, we have created an additional binary column for each category. The only disadvantage of ONE is the number of columns increases immensely. We first one hot encode few columns and then we encode all the columns.



After doing one hot encoding to all columns there will be 145 columns in total.

Now from these 145 columns we have to choose the best features that contribute most to the dataset prediction. For this we use some feature selection functions.



****

****

The top 8 best features are selected using kbest selection method. Now we fit the columns to different models.

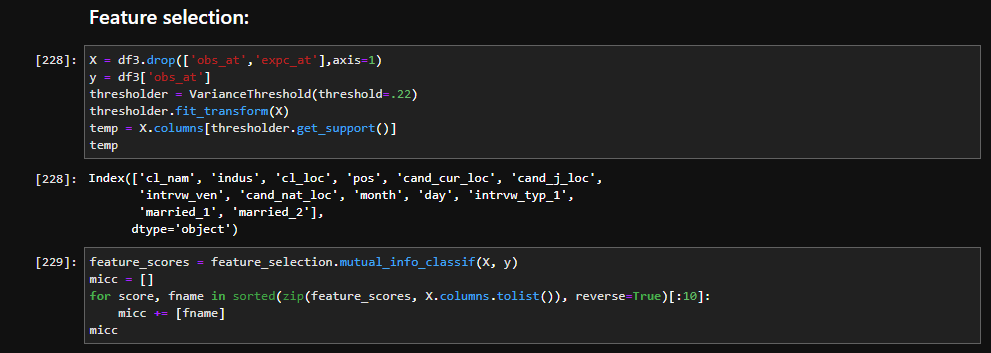
**MODEL BUILDING**

We apply different models:

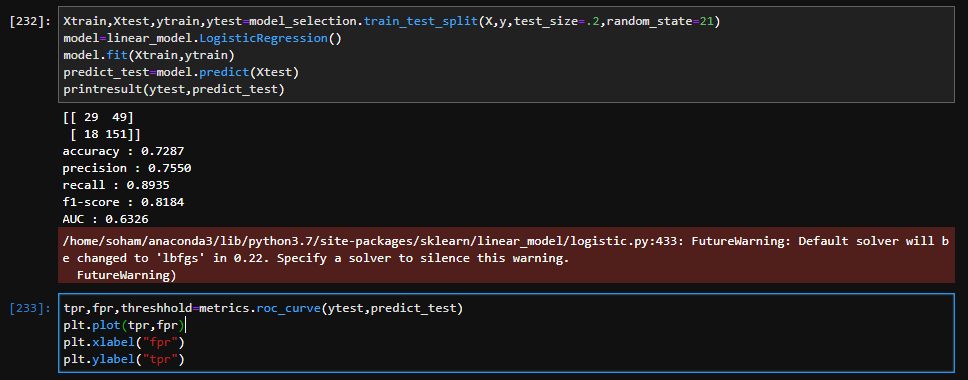
1. Logistic Regression:

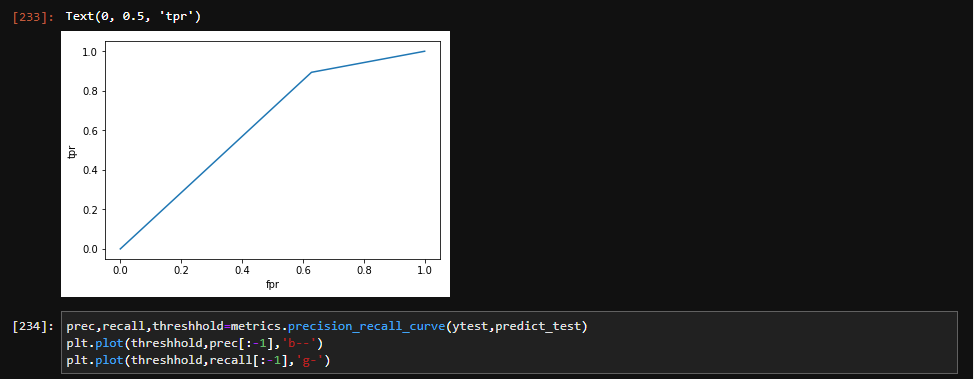
**Logistic regression** is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary dependent variable, although many more complex extensions exist. In **regression** analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a **logistic model** (a form of binary**regression**).

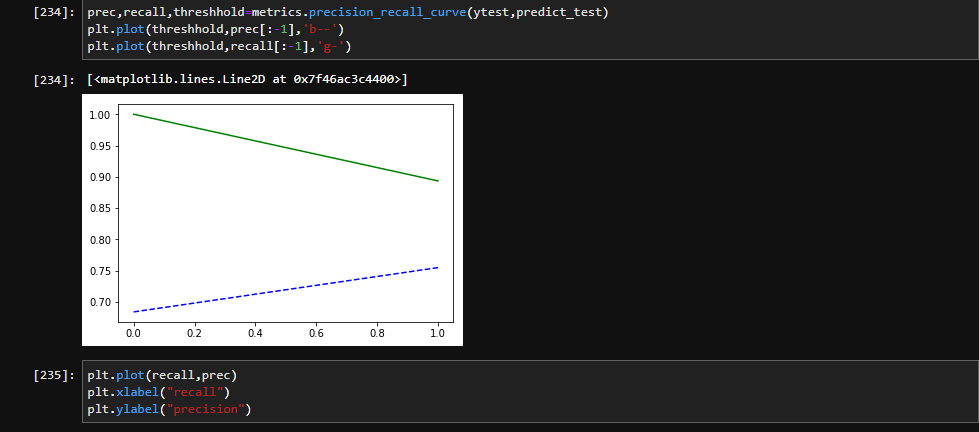
We take one column (obs\_data) for test and rest all columns for training. We use variance threshold for feature selection. We fit Xtrain and ytrain into linear model. Then we predict the result using Xtest. Then we use a user defined function “printresult” to print the confusion matrix and scores of accuracy,precision,recall,f1-score,auc\_roc.

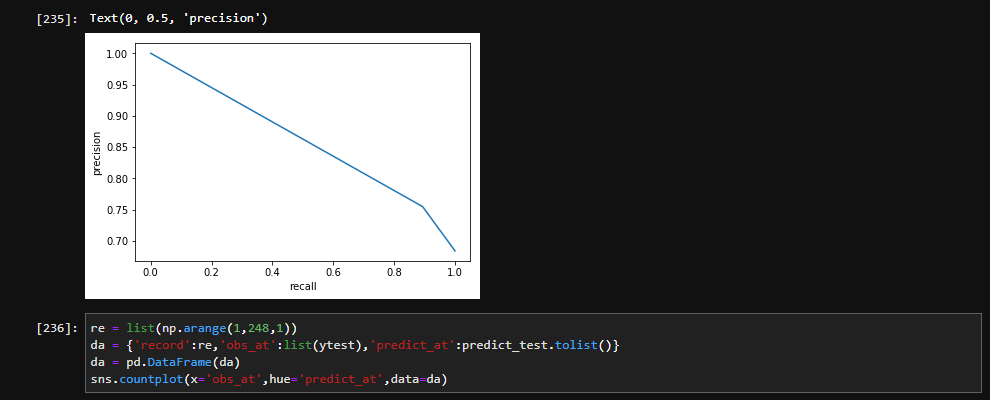




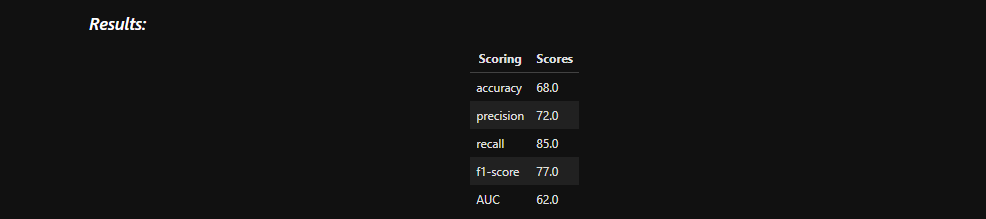












Here after logistic regression model we see precision is 72% and recall is 85%.

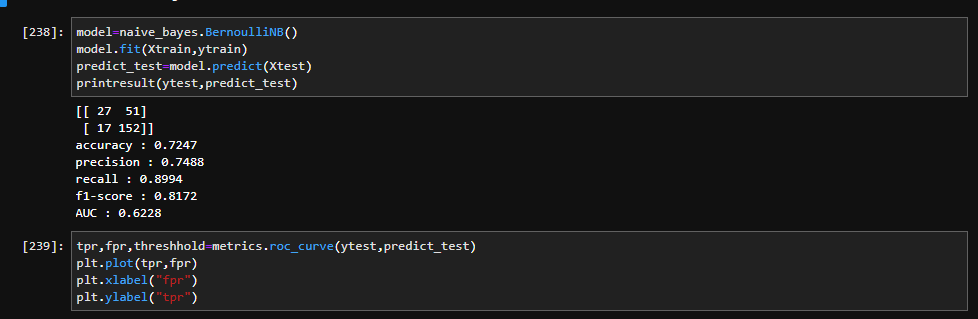
2.Naive Bayes:

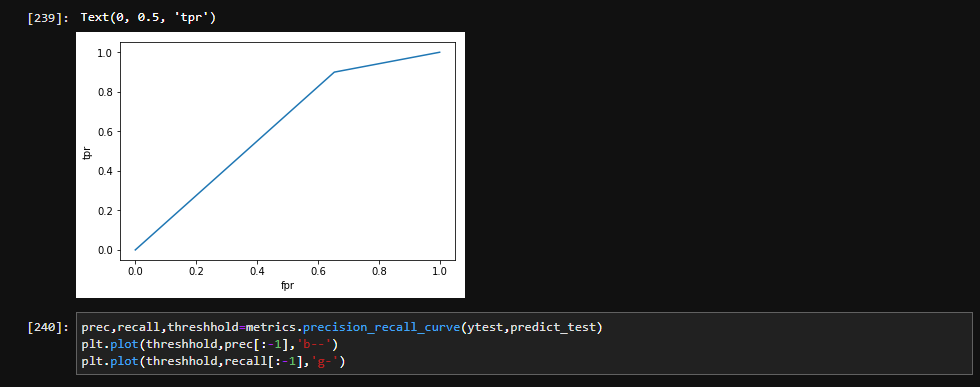
In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

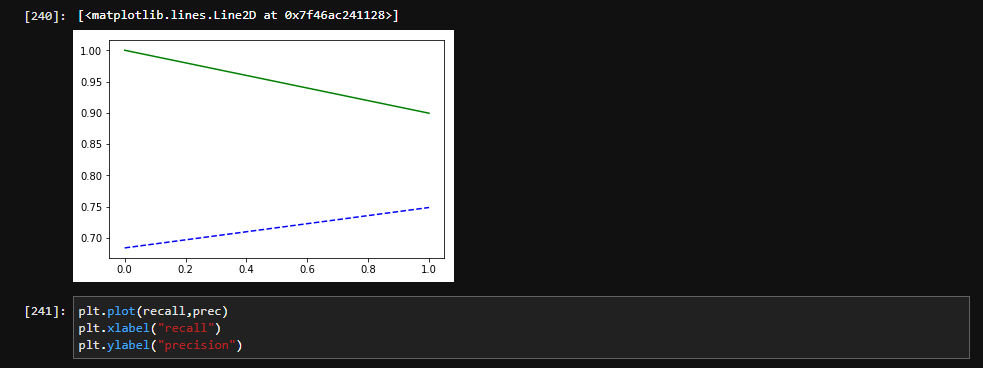
Naive Bayes has been studied extensively since the 1960s. It was introduced (though not under that name) into the text retrieval community in the early 1960s and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines.It also finds application in automatic medical diagnosis.

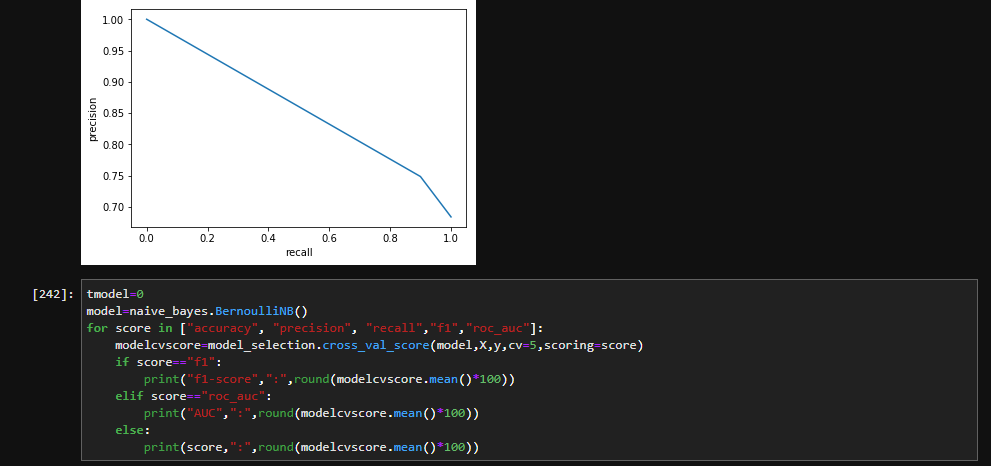
Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression,[4]:718 which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

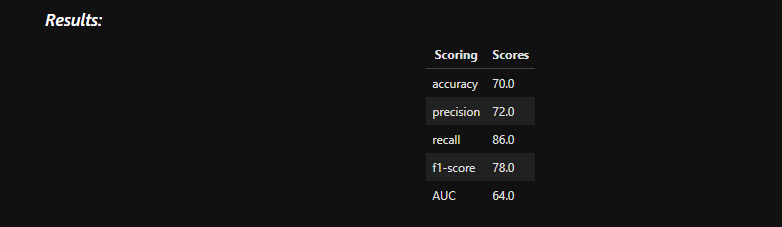
In the statistics and computer science literature, naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method.











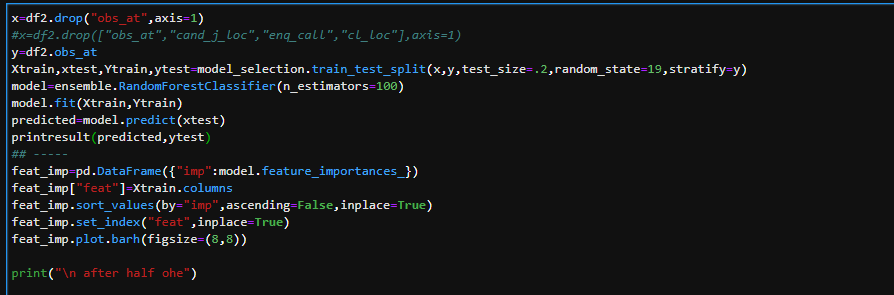
After applying the naive bayes model we see precision is 72% and recall is 86%.

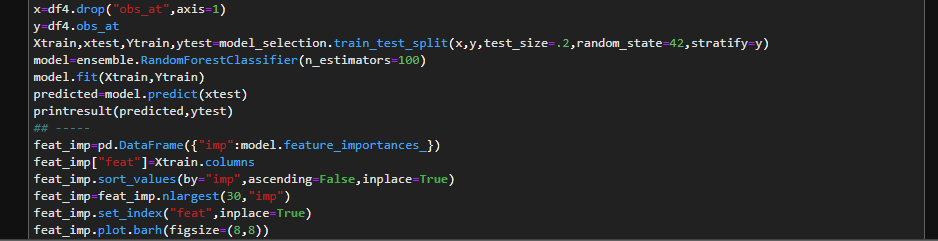
3.Random Forest:

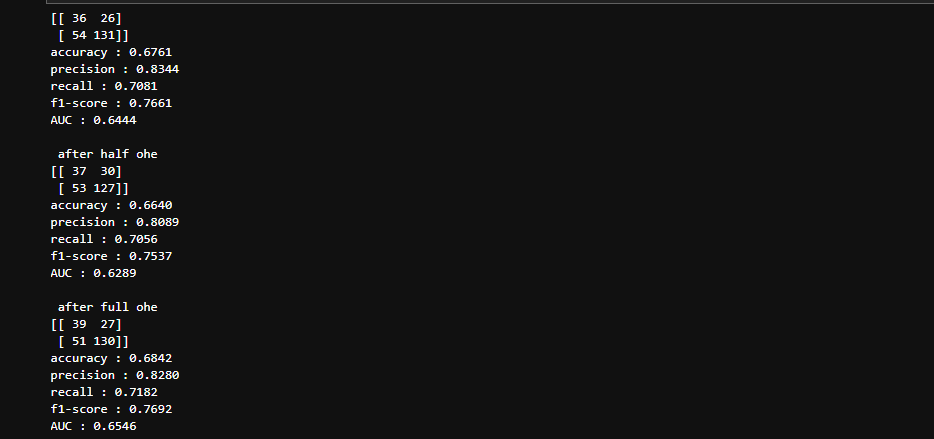
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

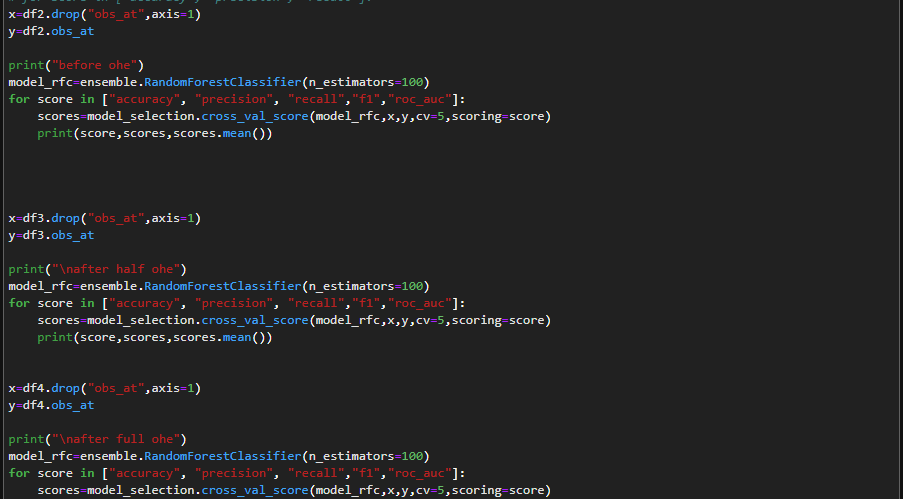
The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

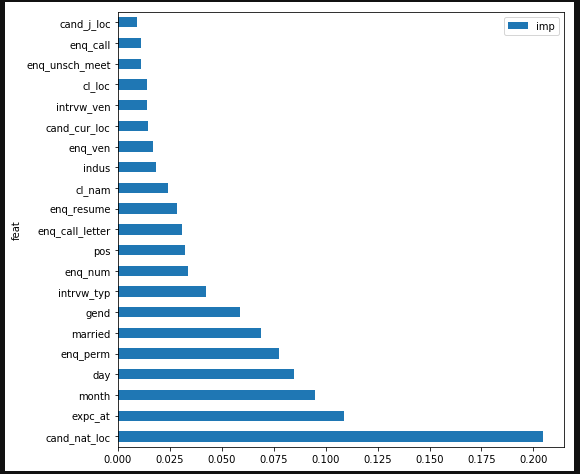
An extension of the algorithm was developed by Leo Breiman and Adele Cutler,who registered "Random Forests" as a trademark (as of 2019, owned by Minitab, Inc.). The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance.

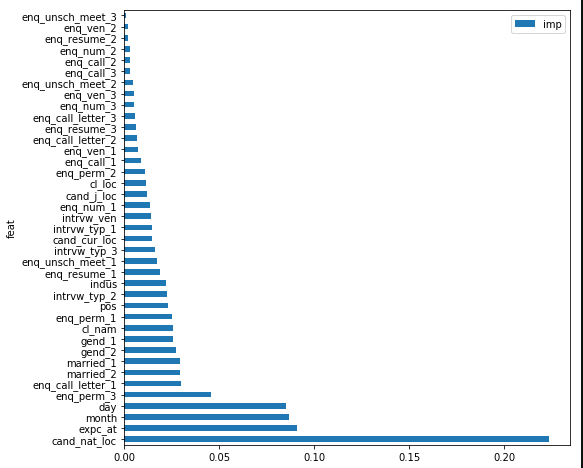


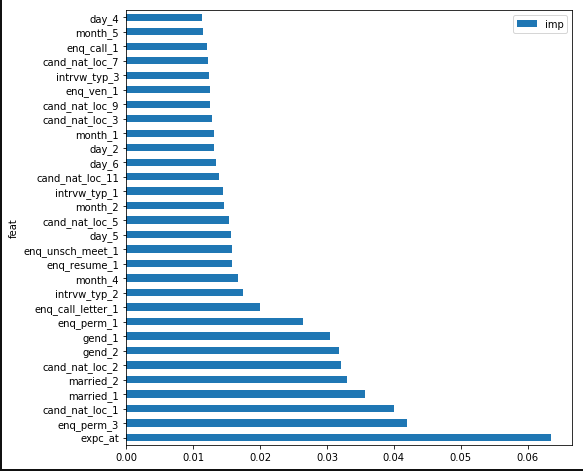


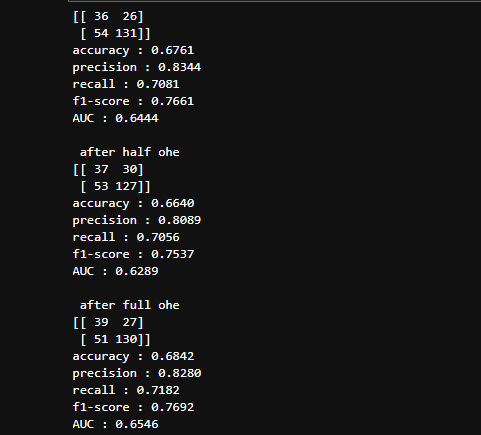








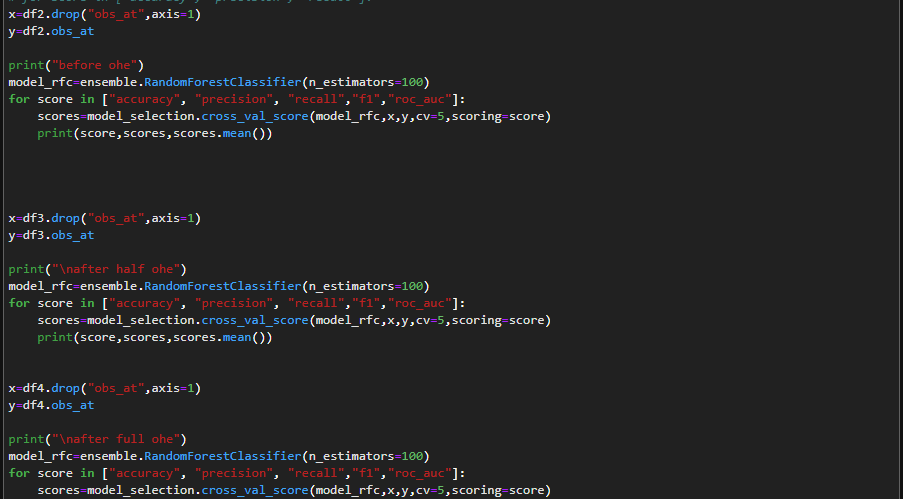


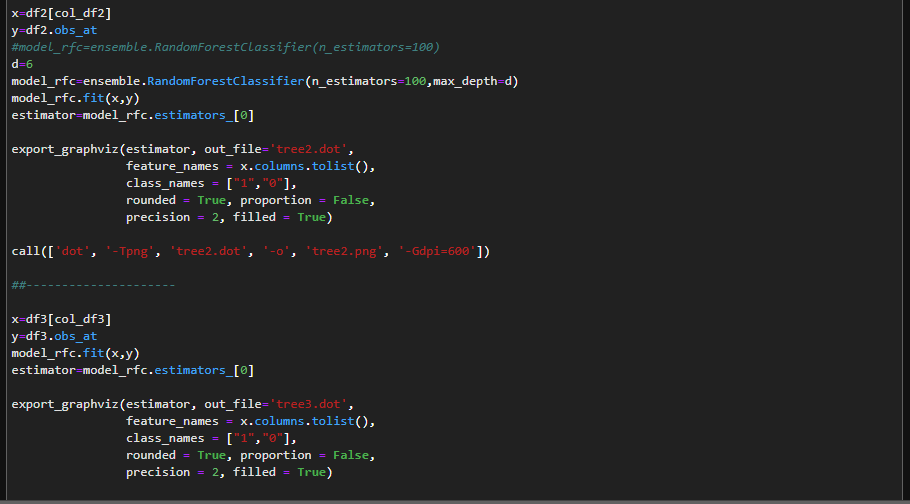


Here now the precision is 82% and recall is 71 %. Now we will crossvalidate the data to see if it overfits or underfits. Cross-validation, sometimes called rotation estimation, or out-of-sample testing is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the validation dataset or testing set).The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

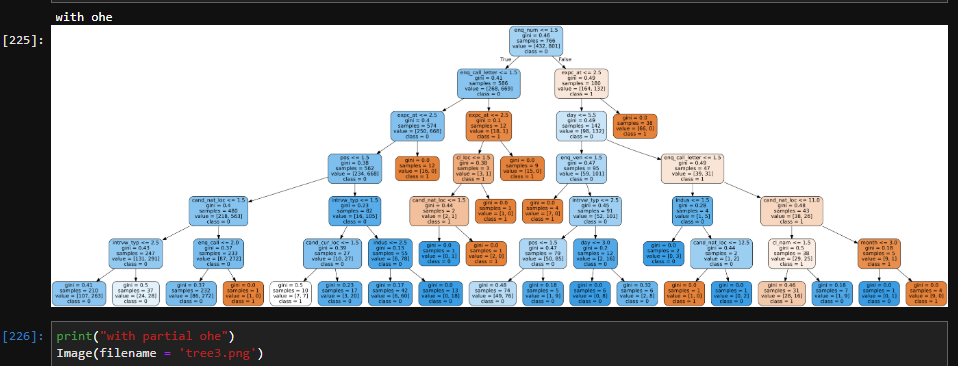
One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, in most methods multiple rounds of cross-validation are performed using different partitions, and the validation results are combined (e.g. averaged) over the rounds to give an estimate of the model's predictive performance.

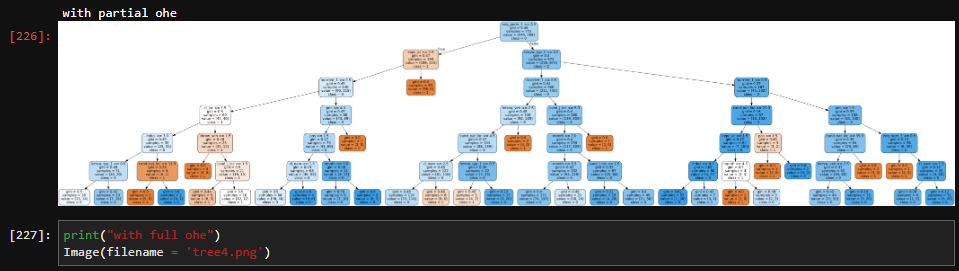
In summary, cross-validation combines (averages) measures of fitness in prediction to derive a more accurate estimate of model prediction performance. Here we use 5 fold cross validation i.e 5 times validation will be checked.

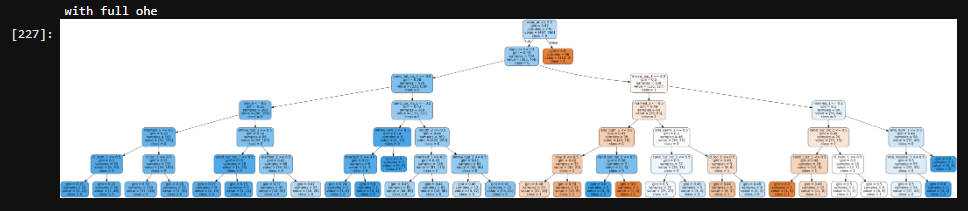


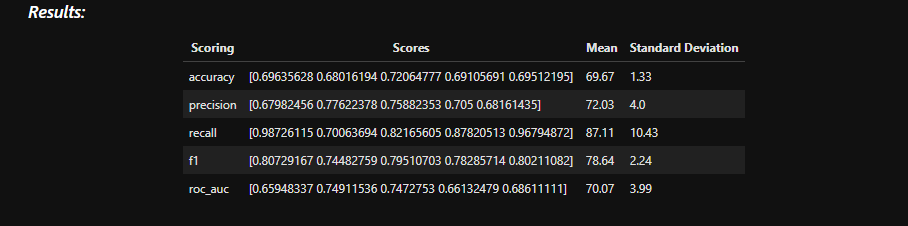












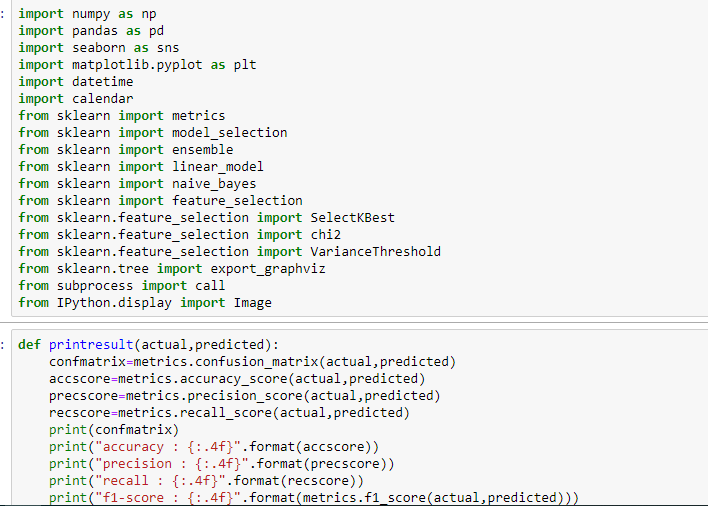
After cross validation we see here precision is 72.03% and recall is 87%.

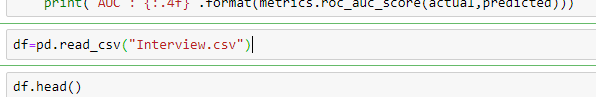
Here after cross validation precision decreased and recall improved. Though the precision decreased the model became more generalized so it’s the best model and it overfits.

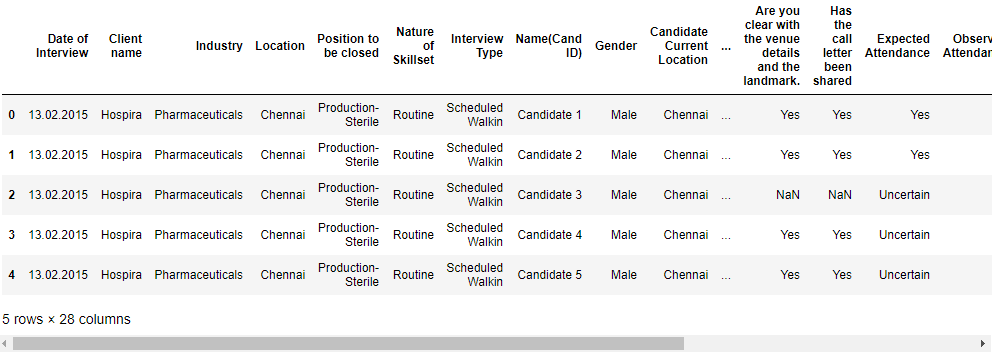
**CONCLUSION**

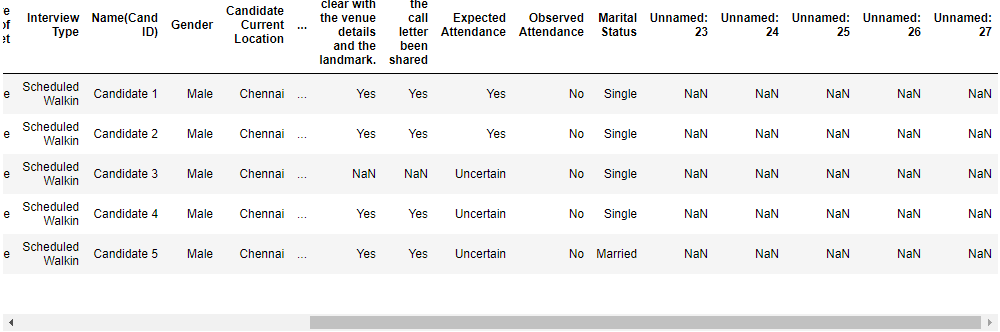
After applying the three models on the cleaned data of the data set we see that random forest is the best method as it gives the highest precision and recall. In random forest method we see after cross validation recall improved and precision decreased so the model overfits.

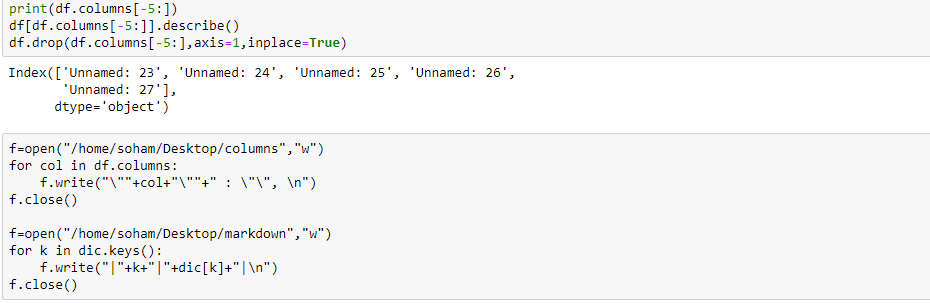
**CODE:**

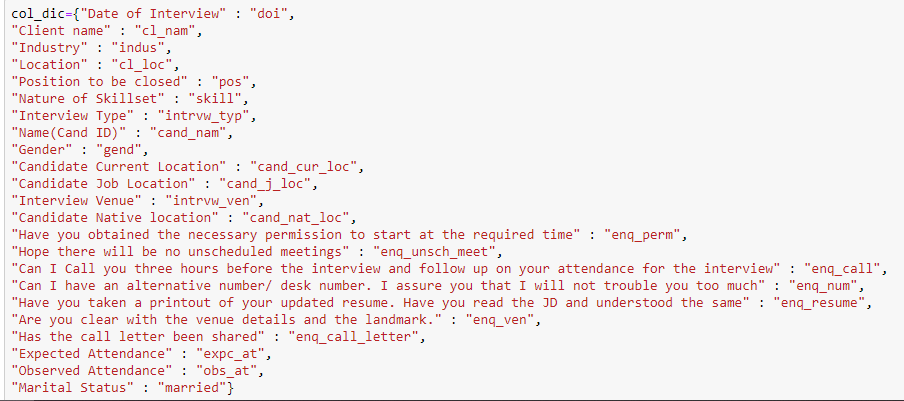




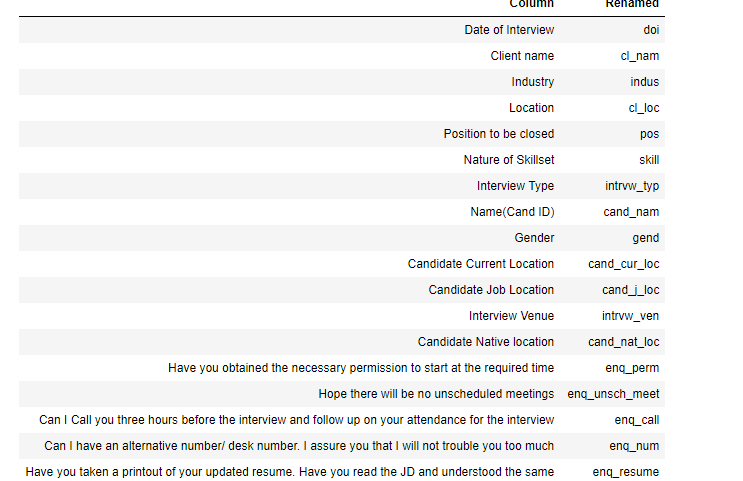


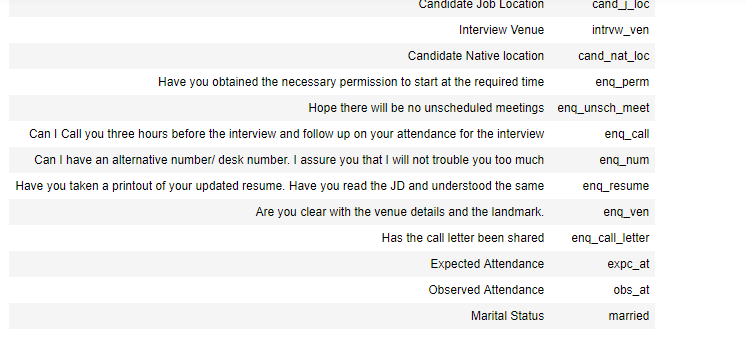


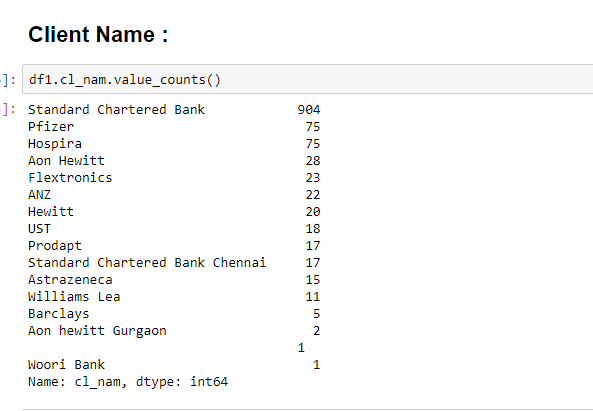


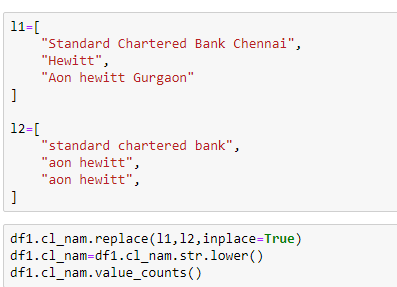


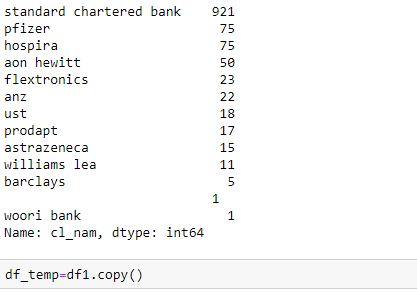


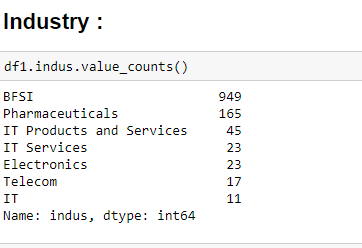


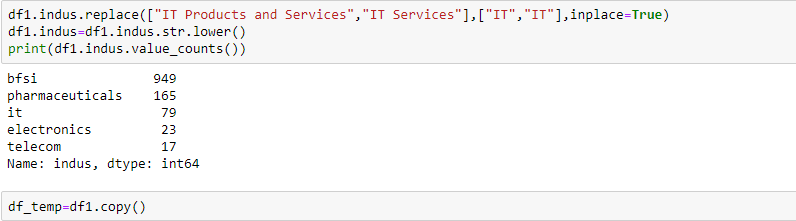




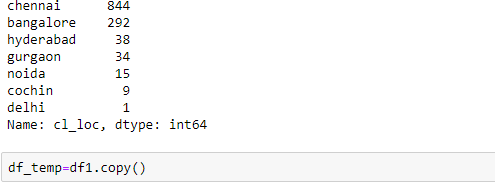


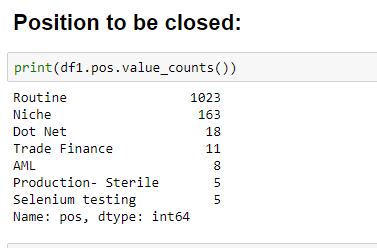


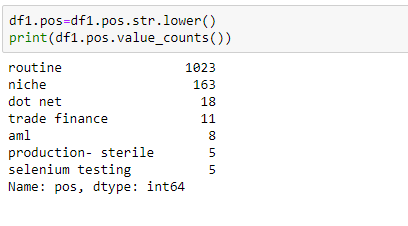


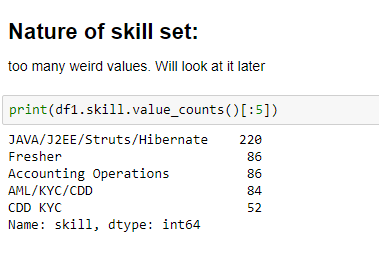




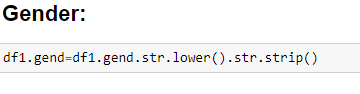


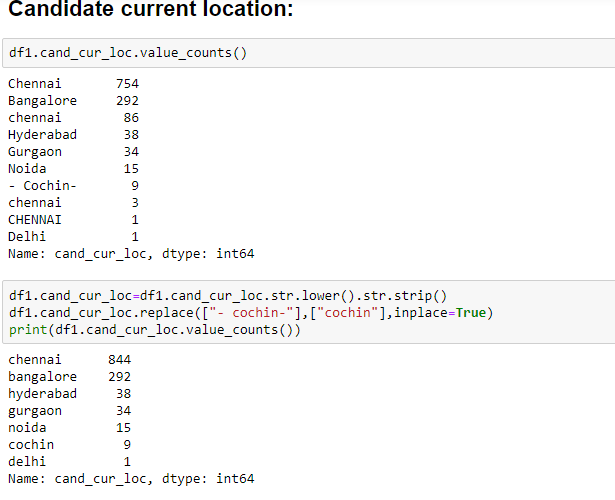


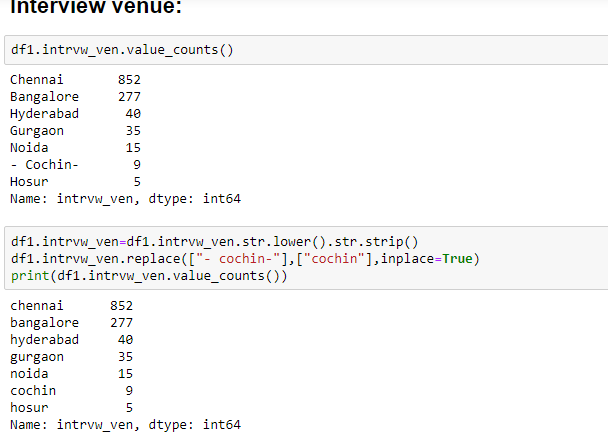




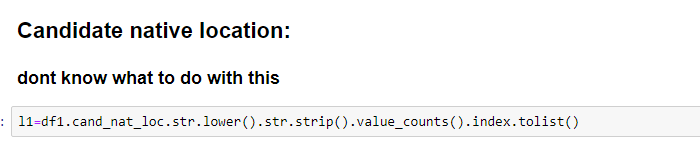




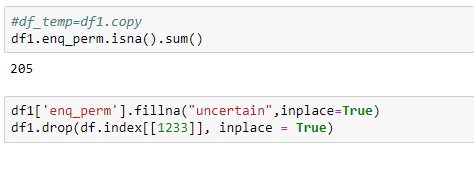


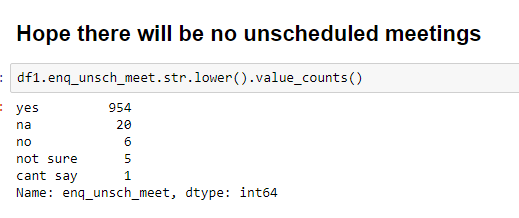


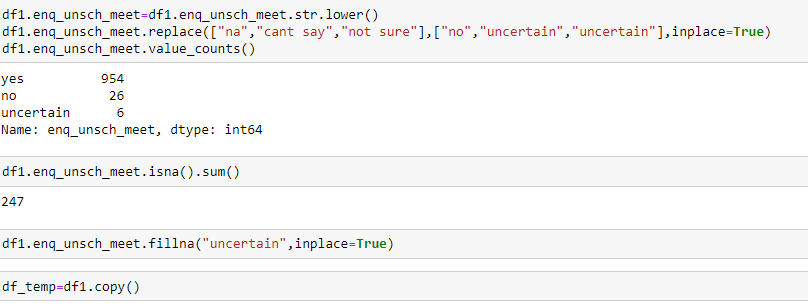




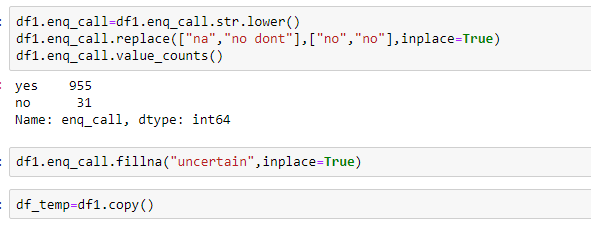


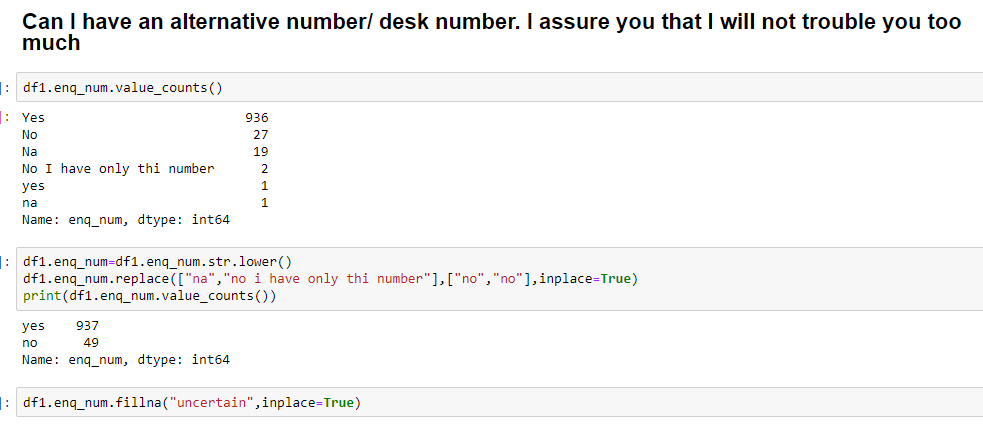




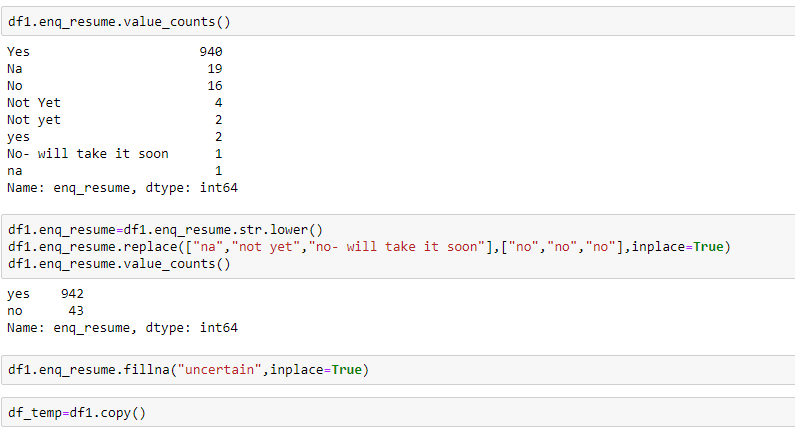


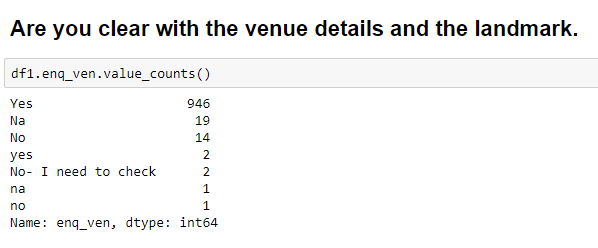


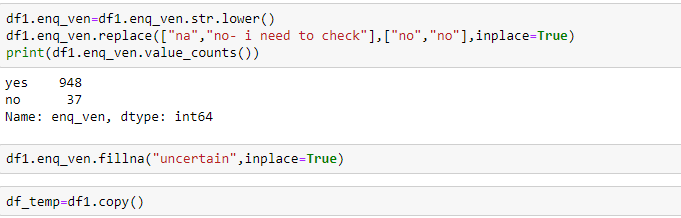


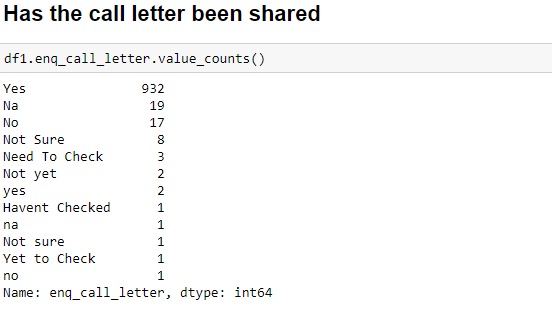


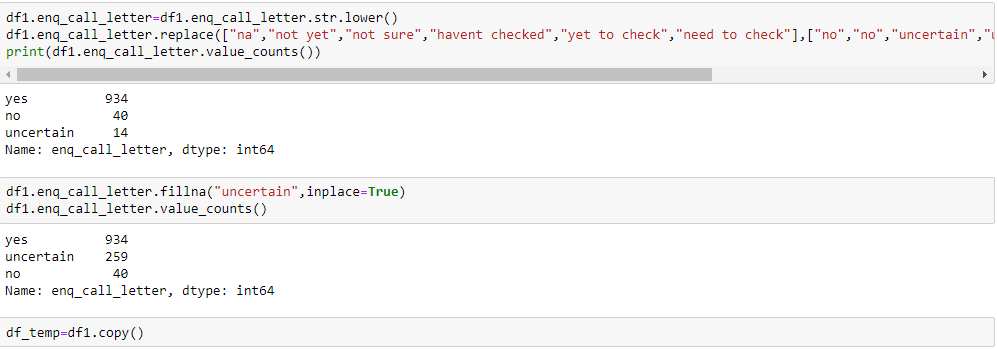


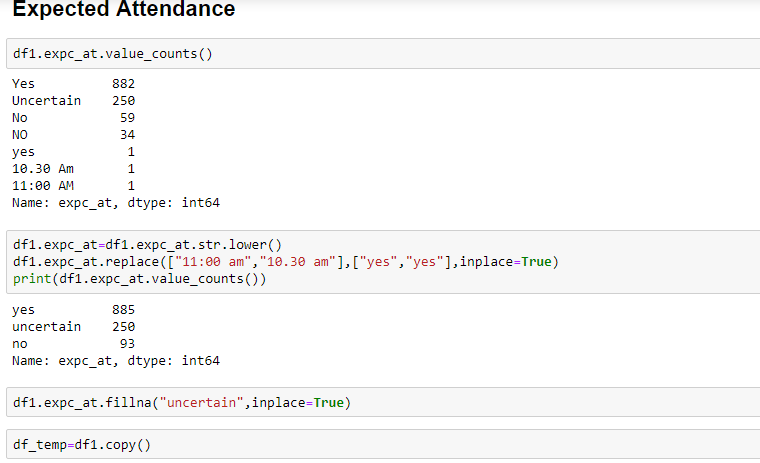


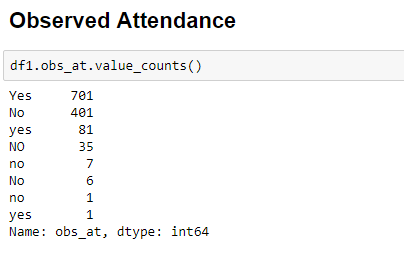


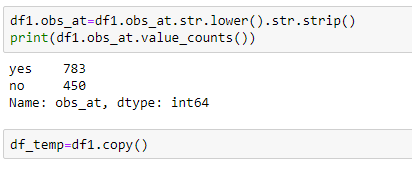


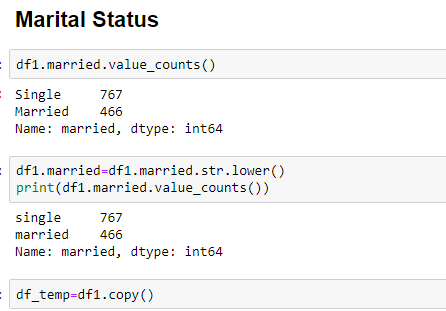


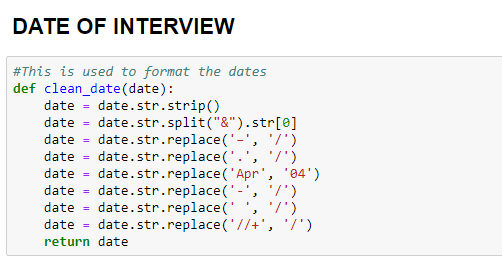


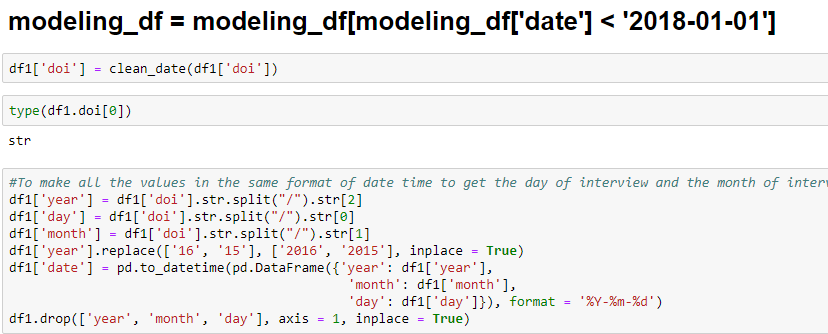




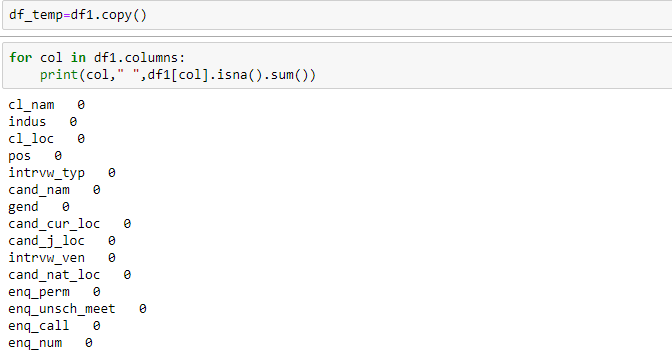


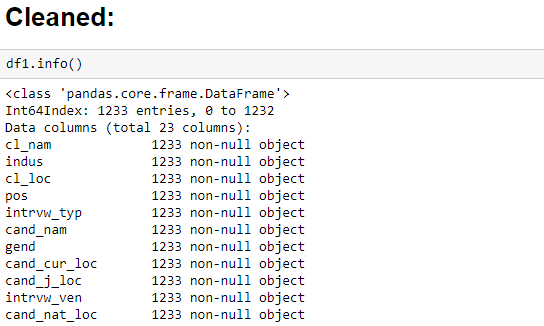


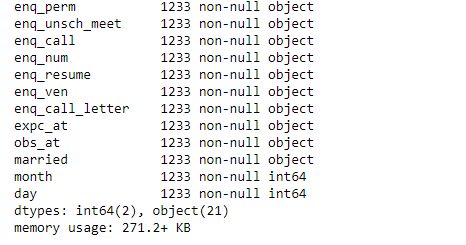




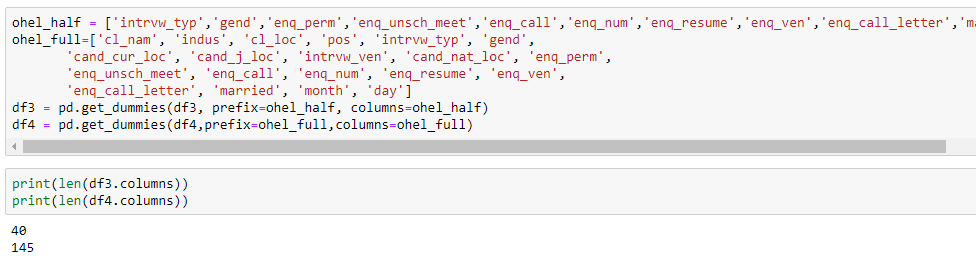


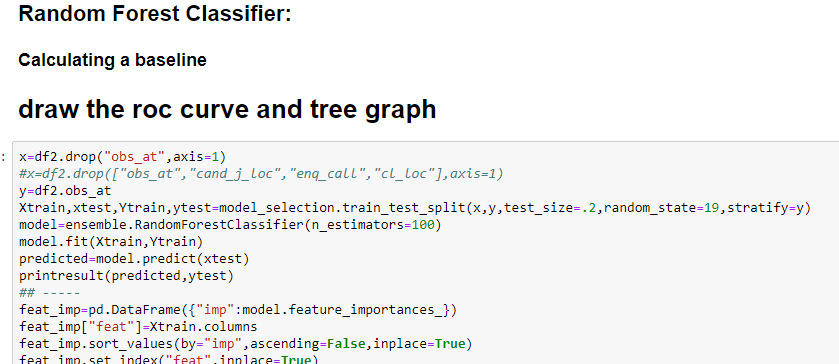


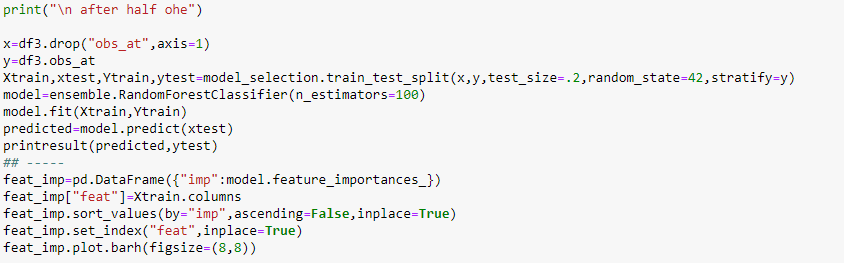


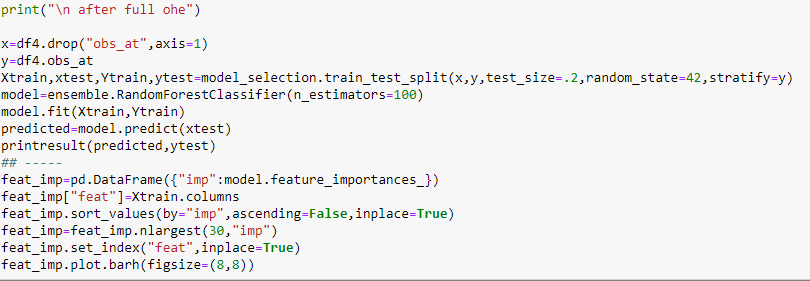




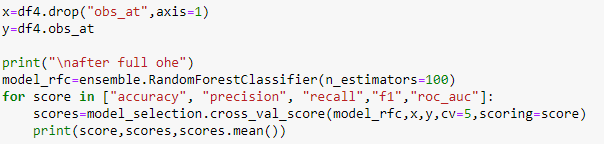




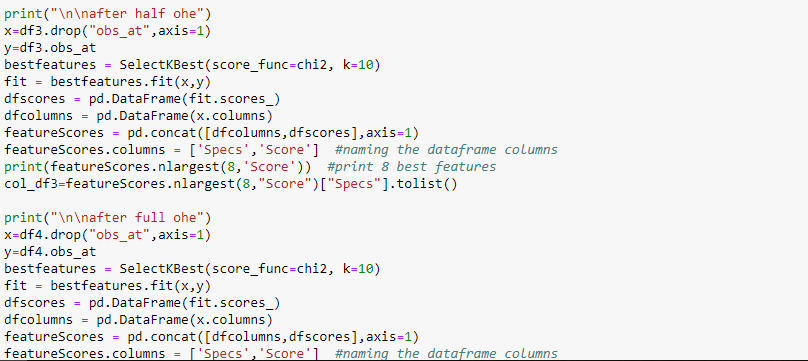


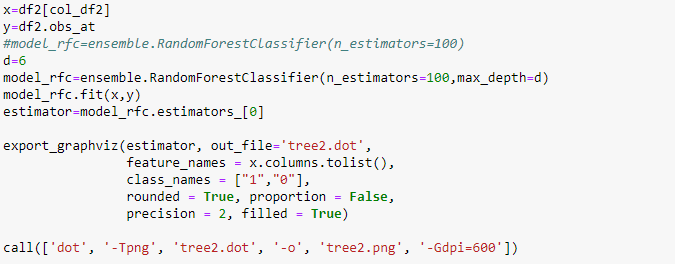


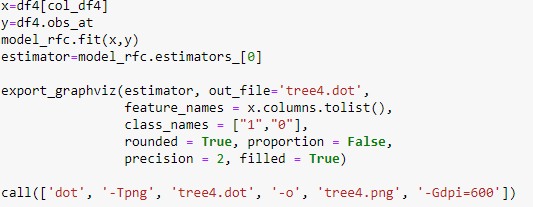




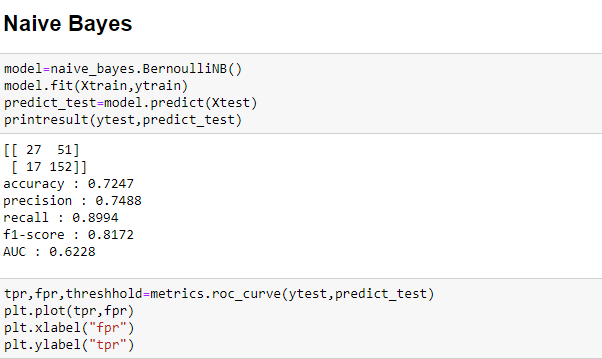


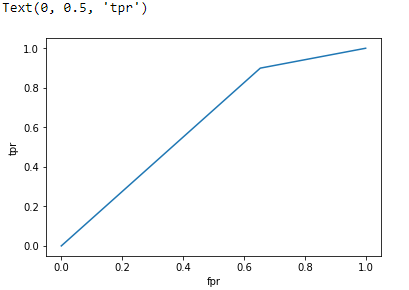


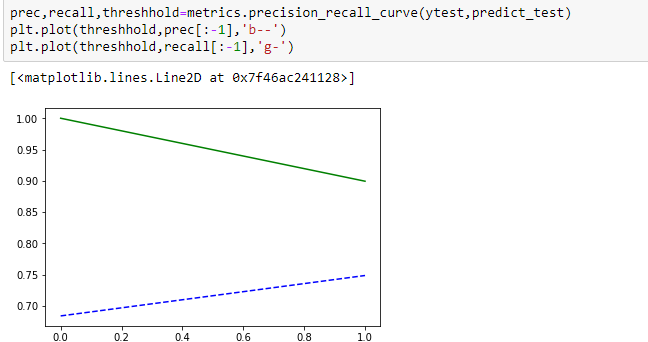


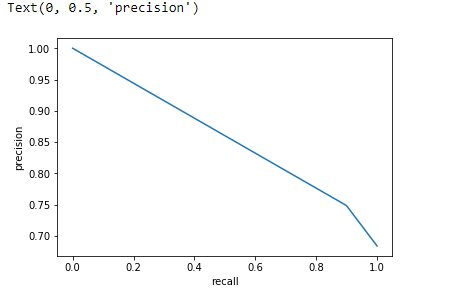














**FUTURE SCOPE OF IMPROVEMENTS**

We have gathered the analysis of data from the data sheet, it can be possible furthermore to predict future. If the information have been more precise then the future predictions would have been more accurate.

CERTIFICATE

This is to certify that Ms. CHANDRABALI BISHNU of RCCIIT,

Registration number: 161170110026, has successfully completed a project on INTERVIEW PREDICTION using MACHINE LEARNING and PYTHON under the guidance of Mr. TITAS ROYCHOWDHURY.

--------------------------------------------------

[*Name of your faculty*]

**Globsyn Finishing School**

CERTIFICATE

This is to certify that Ms. CAMELIA MAHATO of RCCIIT,

Registration number: 161170110025, has successfully completed a project on INTERVIEW PREDICTION using MACHINE LEARNING and PYTHON under the guidance of Mr. TITAS ROYCHOWDHURY.

--------------------------------------------------

[*Name of your faculty*]

**Globsyn Finishing School**

CERTIFICATE

This is to certify that Mr. PALLAB CHAKRABORTY of RCCIIT,

Registration number: 161170110048, has successfully completed a project ON INTERVIEW PREDICTION using MACHINE LEARNING and PYTHON under the guidance of Mr. TITAS ROYCHOWDHURY.

--------------------------------------------------

[*Name of your faculty*]

**Globsyn Finishing School**

CERTIFICATE

This is to certify that Mr. SOHAM MANDAL of RCCIIT,

Registration number: 161170110070, has successfully completed a project on INTERVIEW PREDICTION using MACHINE LEARNING and PYTHON under the guidance of Mr. TITAS ROYCHOWDHURY.

--------------------------------------------------

[*Name of your faculty*]

**Globsyn Finishing School**

CERTIFICATE

This is to certify that Mr. ANIKET CHATTOPADHYAY of RCCIIT,

Registration number: 161170110007, has successfully completed a project on INTERVIEW PREDICTION using MACHINE LEARNING and PYTHON under the guidance of Mr. TITAS ROYCHOWDHURY.

-------------------------------------------------

[*Name of your faculty*]

**Globsyn Finishing School**