**Executive Summary**

Companies are looking to use technology to predict trends and understand behavior for better business performance. Predictive analytics is also used across a wide range of data types to achieve greater value. In this project, predicting if a candidate would attend the interview and investing time on the right candidate would help organizations to save a lot of time during sourcing candidates. For this purpose, predictive models are developed by the use of machine learning algorithms. The project uses a candidate dataset which further goes through the various stages of data analysis. With the use of Logistic Regression and Naïve Bayes algorithms, predictive models are implemented and the algorithm learns to predict new data by getting trained using the candidate dataset. The model evaluation shows that the logistic regression model would be able to predict that the candidate would attend an interview with 72.4% accuracy and the Naïve Bayes model, would predict with 74.1% accuracy. Recommendations on further optimizing and improving the accuracy score of the model is also discussed.

**Chapter** 1

**Introduction**

It has been proven that the success of any organization is largely impacted by the quality of its employees. Research have showed that employers who have been successful in the recruitment process, found that there was a 3.5 times more growth in revenue and twice the profit margin when compared to other employers who weren’t. In order to be successful the employers must decide what type of employers they would need to invest time on and plan how to reach such candidates.

Better recruitment strategies provide improved organizational outcomes during the selection process. Discovering potential candidates to fill the available job vacancies can be a time consuming process. A good recruitment program should invest time on the right candidates for the job. Since recruiting a prospective employee for the job can take a lot of time, it is imperative that the organization create strategies which would help them to invest time for those who are genuinely interested. Taking time to ensure the hiring process will help to recruit the right candidate for the job will save a lot of time and will reduce cost.

Organizations could invest the right amount of time in a candidate by making using of technology. Applying data analytics in the recruitment process can help in forecasting the right candidate for the job. Implementing predictive models within the various talent acquisition stages would pave the way for overall successful operations and growth of organizations.

**Problem Statement**

A lot of time is invested in recruiting the right candidate for a particular job position. For this purpose, the employer will need to go through various stages until they hire a suitable candidate for the job. Among the problems that occurs, finding out if a candidate is really interested in this job position is that which is focused in this project.

**Goal**

Unable to hire the right candidate for the job, has been a major challenge faced by the recruitment team. The demand for getting qualified and knowledgeable people for the job outgrows the supply with so many companies competing to attract the right candidate.

Introducing analytics function into the Human Resource (HR) sourcing process would be a step that could take the organization to the next level to recruit the best talent in the industry. Successful adoption of analytics would depend on understanding the key process metric that affects the outcome desired and helps the organization to make informed decision thus saving them huge cost.

Data analytics can help the organization invest time on the right candidate by making use of data and insights. The objective of the project would be to use the candidate data into insights that drive action. Further, the project also intends to create a model that could help in predicting the outcome by finding out the key metric that would hugely influences the response.

**Chapter 2**

**Review of Literature**

The key to success is in utilizing the data available and understanding how to apply it in the area of the talent-sourcing focus. To achieve this, the candidate data that was acquired for the study purpose had to go through a process of exploratory analysis and then data mining techniques. This large dataset had to be inspected, cleansed and transformed. Finally, the data had to be modelled with the goal of finding information that could help in making conclusions and support decision-making.

The process of data analysis starts by collecting the raw candidate data. The data collection method involves various techniques. Interviews, Questionnaires and Surveys, Observations, Focus groups, Case Studies and Collecting data from documents and records are some of the few techniques that could be used to collect the data for analysis.

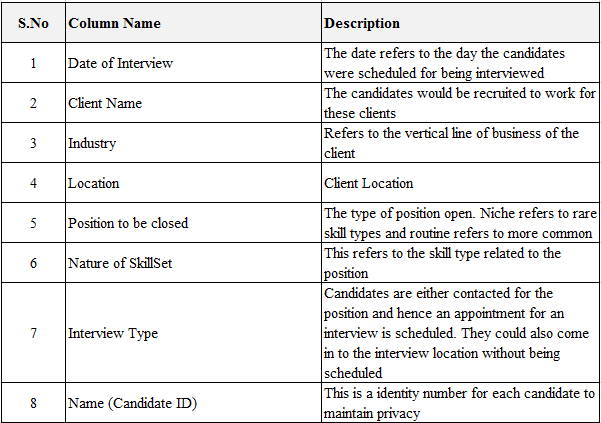
**The Candidate Interview Dataset**

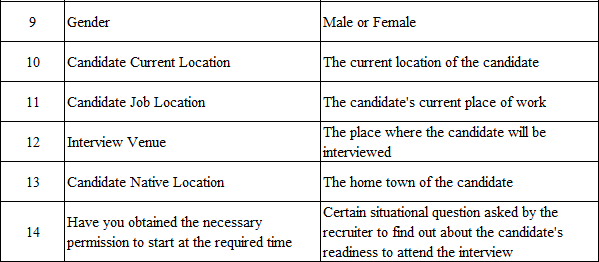
The data used in this project is obtained from a public data platform, kaggle.com, an online community owned by Google, Inc. The dataset that have been used in this project was created by a team of researchers from various HR departments. The information within the dataset pertains to the year 2014 to 2016. The team of researchers who were involved in the data acquisition process are-

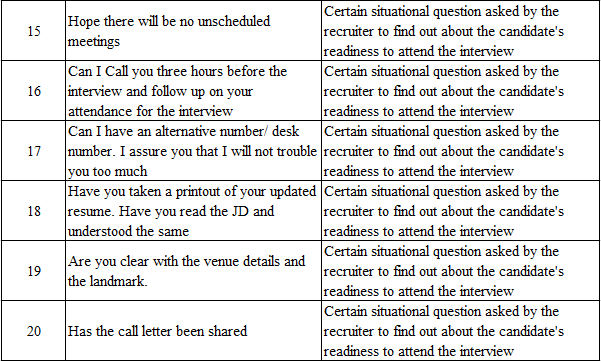
* Dr. Rajendra Desai: Associate Professor, St Joseph's College of Management, Bangalore
* Marcia Akshaya Leo: HR recruiter, Chennai
* Dr. Rashmi Nakra, Professor, St Joseph’s College of Management, Bangalore
* Prima, Student, St Joseph’s College of Management, Bangalore
* Trupthi, Student, St Joseph’s College of Management, Bangalore

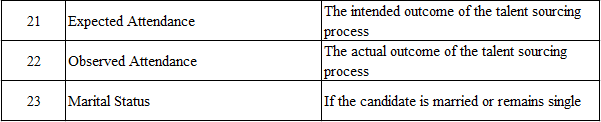
**The Dataset Type**

The candidate interview dataset contains 1234 records and the information within the dataset is related to the candidate and the interview process, refer Appendix A.



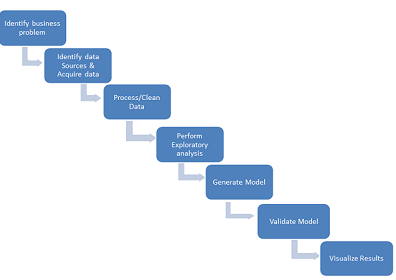






**Data Analysis**

The steps for solving a data analysis problem is shown as below –



**Data Cleansing**

After the candidate dataset acquisition process, the data have to be cleaned. This is considered as one of the important steps in the data analysis process. The quality of the data will very much impact the accuracy of the analysis. There are a few approaches that could be considered here, that is, formatting the data depending on the data analytical tools that would be considered. The missing data should be handled. Data transformations like identifying the outliers would be another approach that could be considered.

**Exploratory Analysis**

To understand the main characteristic, the data has to be explored visually. The analysis is done visually using tools. Exploratory analysis, help us to understand the causes of an observed event. It also helps to understand the nature of the data that is dealt and assumptions are assessed on which the analysis would be based. The data also have to be analyzed for identifying the key features in the analysis.

**Different Techniques Used**

The different techniques that could be used are graphical techniques like scatter plots, box plots, histograms and quantitative techniques like mean, mode, median, standard deviation.

**Model Generation and Validation**

The steps involved in this stage would be model selection, model training and model evaluation. Based on the type of business problem that is dealt with, a model will be built. In case the objective is to predict a future event, then a Regression model could be built. After the model is selected for analysis, the entire dataset could be split for training and testing. Further the model is tested and evaluated. The dataset that is used for testing the model would be (one-third) of the dataset that was split in the previous step.

**Data Visualization:**

At this step, the results obtained from the model and the problem solved would be generally be presented to the higher management for decision-making.

**Chapter 3**

**Methods and Procedures**

To address the problem statement, historical past data was analyzed to forecast the possibility of a candidate to attending an interview. Advanced algorithms that used statistical techniques were applied to the candidate dataset to weigh different variables. These variables were then scored to predict the likelihood that a candidate would attend an interview. Advanced tools and programming languages were used to accomplish the objective.

**Tools and Technology used**

The programming language that is used in this project to draw value based insights is Python. Python is a general purpose programming language that is easy to use, when it comes to analytical and quantitative computing. Python has massive libraries for manipulating data. Libraries like ‘numpy’ and ‘pandas’ are the ones that is used in this project for exploring and manipulating data. ‘Matplotlib’ and ‘seaborn’ are other libraries that is used for developing data visualizations. The machine learning library that is used in this project to develop the predictive model is ‘scikit-learn’. The purpose of machine learning algorithms is to automate the data analysis task on a large scale, on massive volumes of structured or unstructured data. Also, the data is explored within an integrated environment like ‘Jupyter Notebook’. This is an open-source web application which allows to develop code for manipulating data and exploring data using visualizations. The environment also allows to share and save code for interactive computing.

**Method or techniques used**

Through this project we aim to use algorithms to help us predict if a candidate is interested or not. To achieve this, we would be using supervised learning algorithms, a type of machine learning algorithm, to teach the computer on how to achieve a result. To explain further, we would be providing data to the learning algorithm to which we know the right answers and then the algorithm would learn from it and produce more of the right answers.

To determine if a candidate is interested in the interview or not would be a classification problem. That is here we try to predict a discrete value (0 or 1) output with various attributes or features. The specialized algorithms that we plan to use to create predictive models are:

**Logistic Regression model**

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable. The technique could be used for finding the cause and effect relationship. It is also used because of its benefits like:

1. It indicates the significant relationship between dependent variable and independent variable.

2. It indicates the strength of impact of multiple independent variable on a dependent variable.

One of the regression techniques available to make prediction is the logistic regression model. This predictive analysis technique is appropriate method for binary classification problems, that is when the dependent variable is binary. It is a simple algorithm which is easy to implement and will do well in many tasks. The dependent variable that we use in our project is the ‘observed attendance’ which has values ‘Yes’ or ‘No’.

**Naïve Bayes:**

Naïve bayes algorithm, a type of machine learning algorithm is one of the supervised learning approach used to model a predictive modeling problem probabilistically. Naïve bayes algorithm is an intuitive method which uses the probabilities of all the attributes to make desired predictions. Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. Naïve Bayes is a strong assumption method but results in a fast and effective method and generates robust results. Naive Bayes can be used for the prediction even by using a small amount of training data.

In our project, the candidate’s presence to the interview has to be predicted. With the available parameters, we can use the algorithm to predict whether the candidate will attend the interview or not. Naïve bayes is an extremely fast algorithm when compared with the other algorithms.

In the available dataset we have both the inputs and the outputs for the project. We intend to use Naïve Bayes algorithm to do the prediction. Once the dataset and the algorithm are fed, it uses the probabilistic model which uses the probabilities of each attribute belonging to each class to make a prediction. Based on the available dataset, to make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

**Procedures followed**

Step 1: The interview candidate dataset was imported into the tool for data cleansing, exploration and visualization.

Step 2: The response variable, Observed attendance, and other categorical data were transformed into 0s and 1s for statistical analysis.

Step 3: Implementing machine learning algorithms like Logistic Regression and Naïve Bayes to develop a predicting model.

Step 4: Splitting the dataset to be used for training and testing the model

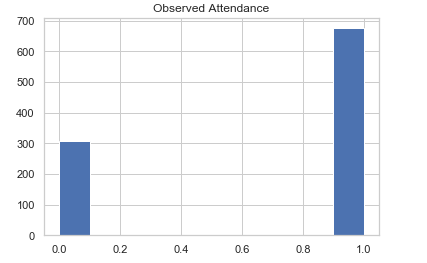
Step 5: Calculating the accuracy of the developed model to check how precisely the model will work.

**Chapter 4**

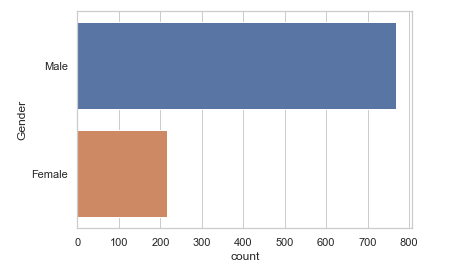
**Results**

The project has addressed the binary classification problem of predicting whether the candidate will attend the interview or not. The dataset was cleansed before being further explored. After removing the duplicate values and the blank values, the variables within the dataset were visually represented.

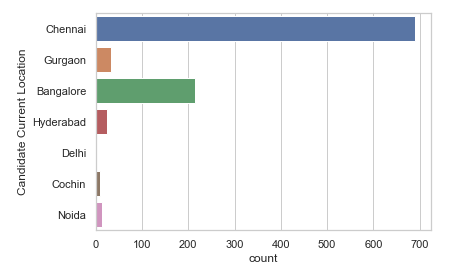
Observed Attendance



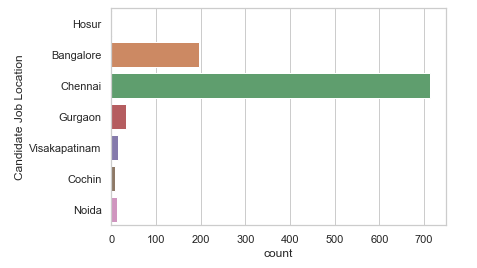
Gender



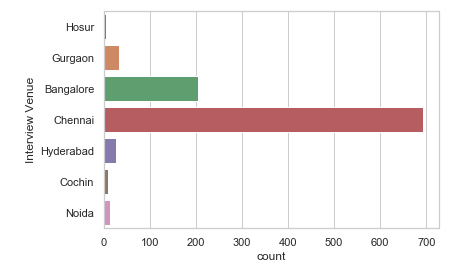
Candidate Current Location



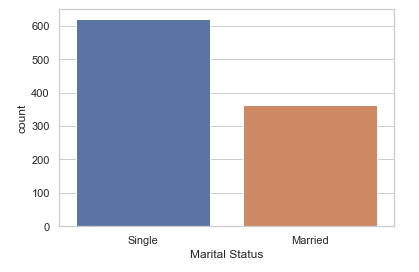
Candidate Job Location



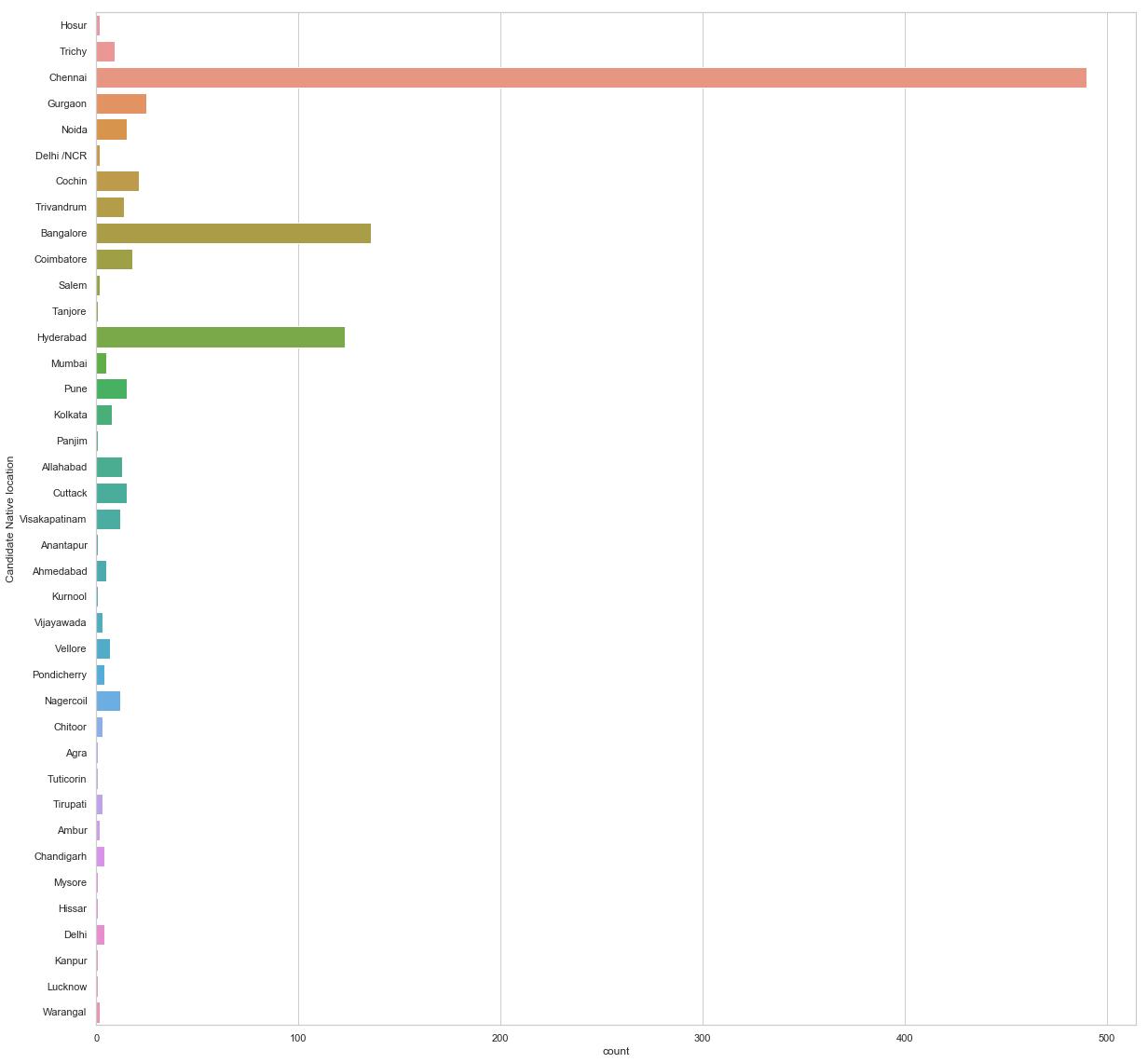
Interview Venue



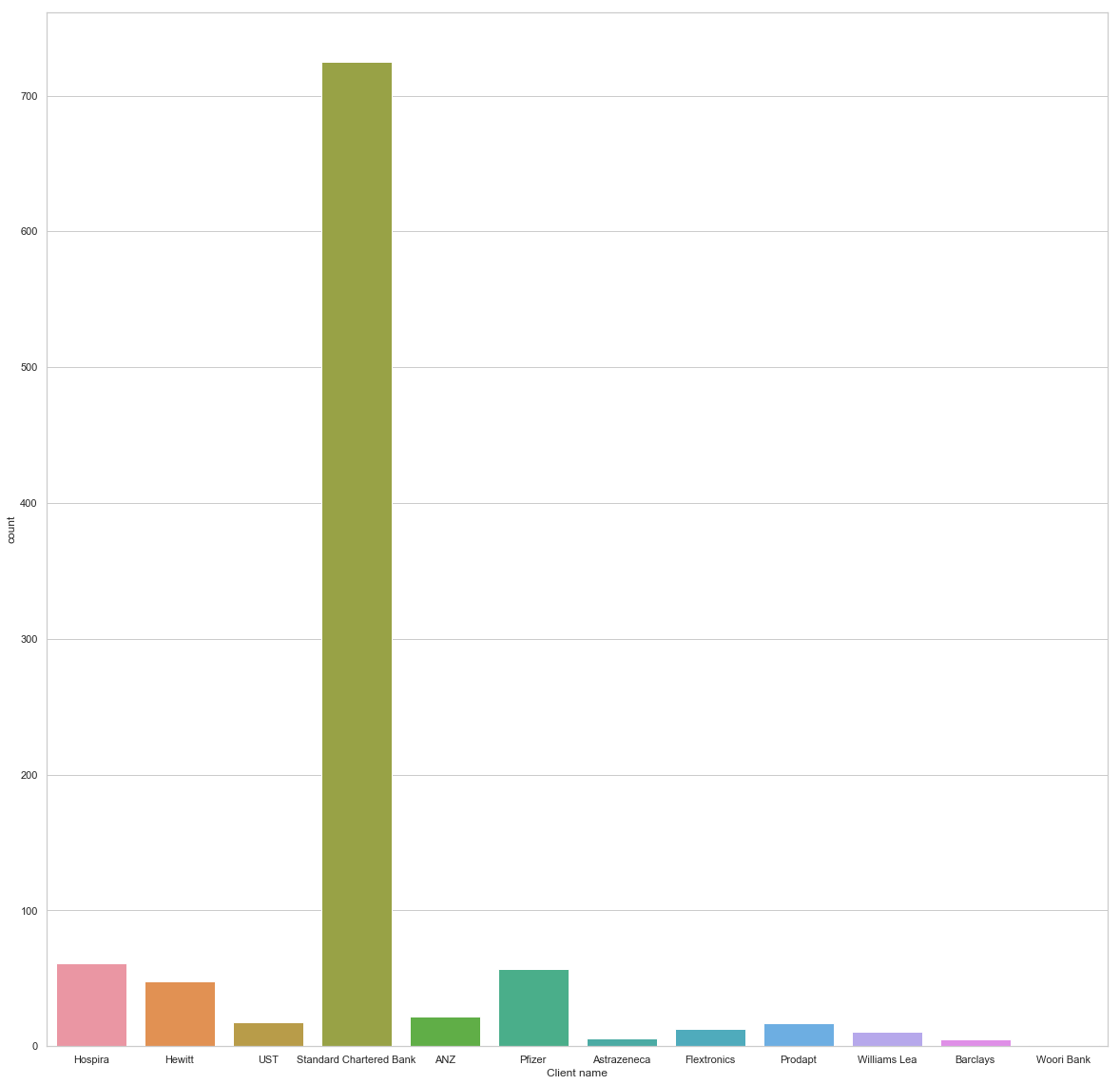
Marital Status



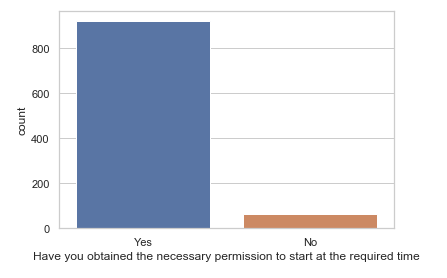
Candidate Native Location



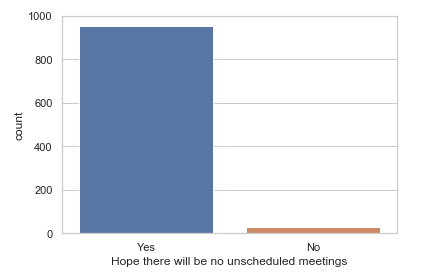
Client Name



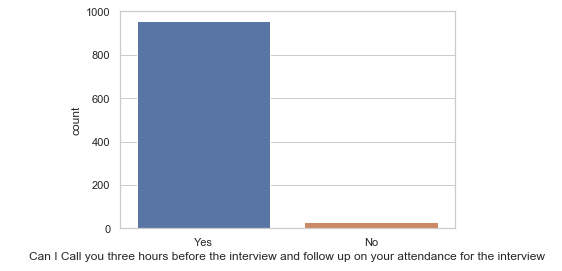
Have you obtained the necessary permission to start at the required time?



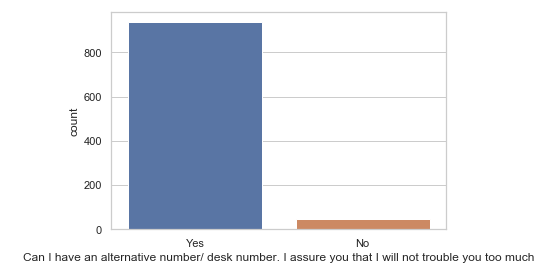
Hope there will be no unscheduled meetings?



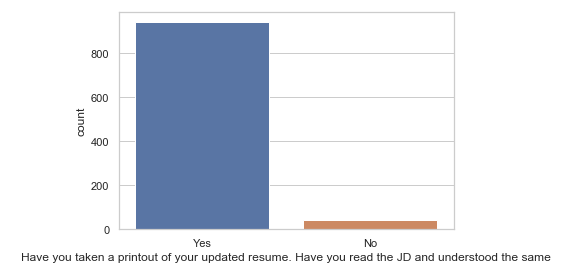
Can I call you three hours before the interview and follow up on your attendance for the interview?



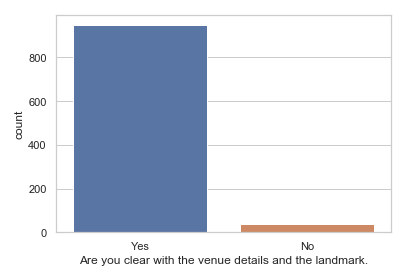
Can I have an alternative number or desk number. I assure you that I will not trouble you too much?



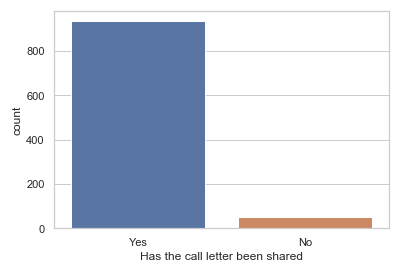
Have you taken a printout of your updated resume? Have you read the JD and understood the same?



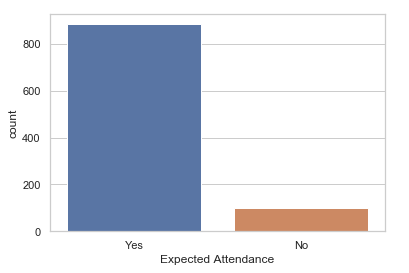
Are you clear with the venue details and the landmark?



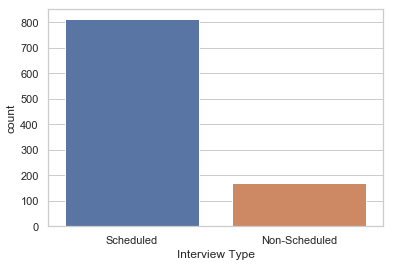
Has the call letter been shared?



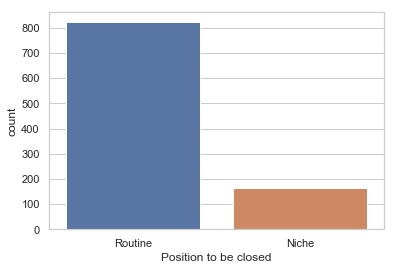
Expected Attendance



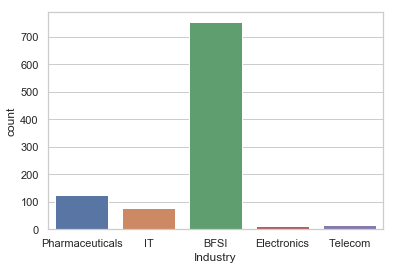
Interview Type



Position to be closed



Industry

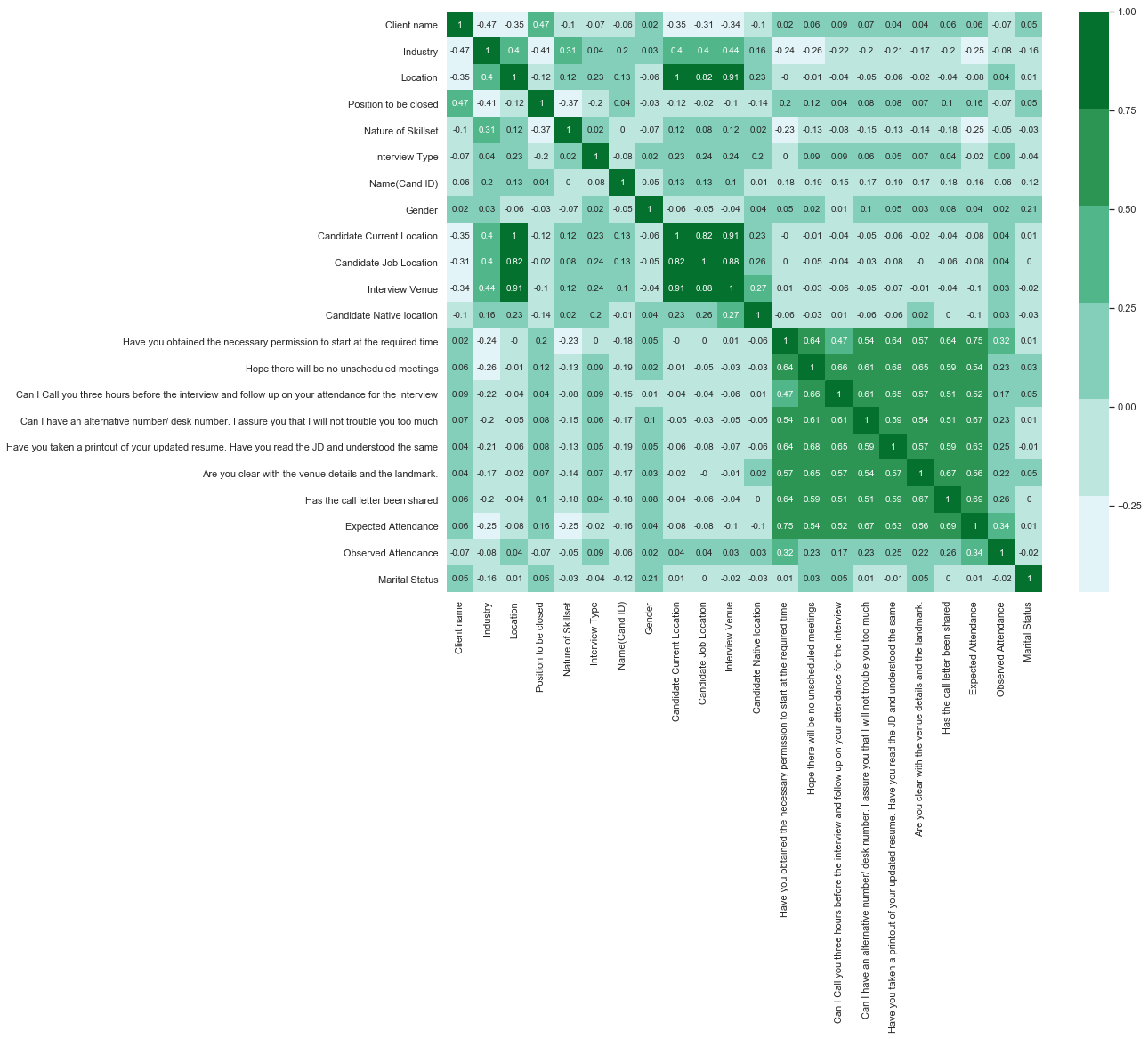


The binary target variable used in this project is the Observed Attendance variable. The observed attendance variable has two values, that is, Yes (that candidate attended the interview) or No (that candidate did not attend the interview).

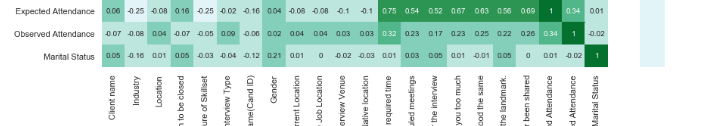
**Selecting the features**

Feature selection aims to use the most relevant features to the problem and removes the unnecessary features. The unnecessary features include redundant and noisy features. The purpose of this is to reduce the number of features and reduce over fitting. Feature selection also aims to improve the generalization of the model. It is also used to understand the features and their relationship to the target variable. The process could also involve the use of the learning algorithms. There are several methods to select the relevant features. Mutual information, correlation matrix, chi-square are some of them which could be used to find the dependency between the variables.

The simplest method to understand feature selection to the response variable is Pearson’s Correlation Matrix and is used in this project. The resulting value lies between 1 and -1. Obtaining 1 means there is a positive correlation, -1 indicates a negative correlation and 0 would mean no correlation.



The values obtained to determine the strength between the variables and the response variable, observed attendance is shown below. The picture is extracted from the above picture for a clearer view.



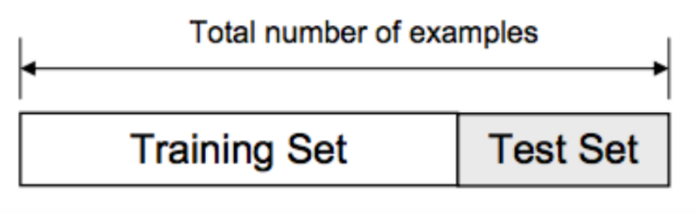
An observed value can occur for reasons like sampling variation or because there is a genuine association between the variables.

**Splitting the data into train and test data**

Next the dataset is split into training data and test data. This is done to fit our model on the train data and make predictions on the test data. During this process, one of the two things is expected to happen. During the process, our model could either be found to overfit or underfit. This has to be avoided since it would affect the predictability of the model. A model is overfit when it has been trained too well and fits very closely to the dataset. This model would be very accurate on the training data but could predict inaccurately on a new data. This is because the model turns out to be as ‘not-generalized’. This means that the model learns or describes the noise in the training data and misses out on the actual relationships between variables. On the other hand, underfit means, the model does not fit the training data and therefore misses on the trends in the data. This could happen because of not enough predictors or independent variables in the data.



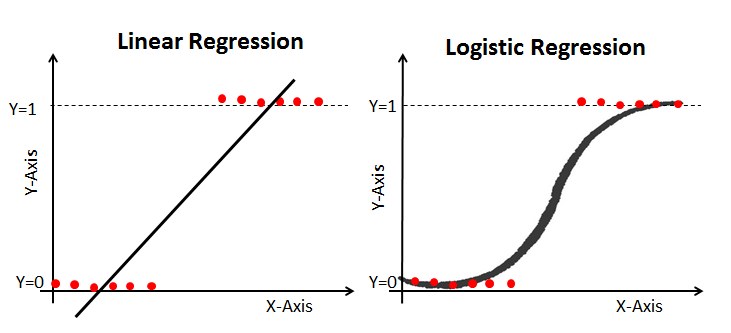
As mentioned earlier, the dataset is split into training and test data. The training dataset contains a known output and the model learns on this data inorder to be generalized on other data in the future. The test data is used to test the model’s prediction on the subset.



The scikit-learn library is used to perform the test-train split and the dataset split is usually around 80/20 or 70/30. The cross validation library is used to reserve a particular sample of the dataset to test the model. The test-train split function uses a random-state variable which is set to a particular number to guarantee that the same sequence of random numbers is generated each time the code is run. The results produced will always be the same unless there is other randomness present in the process. This helps in verifying the output.

**Implementation using Logistic Regression**

The outcome or target variable has only two possible values. It is a special case of linear regression where the target variable is categorical in nature. The dependent variable is a log of odds and a logit function is used to predict the probability of occurrence of a binary event. Linear regression gives a continuous output but a logistic regression provides a constant output like if a candidate will attend the interview or not. Logistic regression is estimated using Maximum Likelihood Estimation approach.

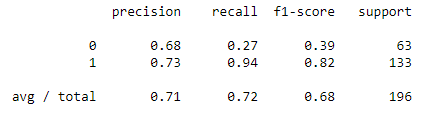


The sigmoid function, also called logistic function gives the S shaped curve that could take any real-valued number and map it into a value between 0 and 1.

f(x) = 1/(1+e^(-x))

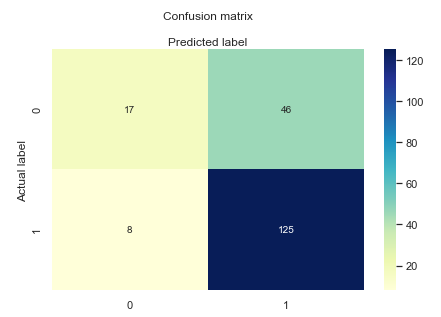
A logistic regression classifier instance is used which is then used to learn from the model. This is done by passing the training set to the fit method. Inorder to predict the new values, we would then pass the test dataset or the new data to the predict method.

The sklearn metrics module implements loss, score and utility functions that helps to measure classification performance. The classification metrics, precision tells the ability of the classifier not to label as positive, a sample that is negative. The recall metric gives the ability of the classifier to find all positive samples. The F score gives a weighted harmonic mean of the precision and recall. The F score reaches its best value at 1 and worst score at 0. The support is the number of occurrences of each class in y\_true.



Confusion Matrix

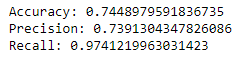
This is a table that is used to evaluate the performance of a classification model. The aim of the confusion matrix is to determine the number of correct and incorrect predictions that is summed up class-wise. The diagonal elements give actual predictions and non-diagonal elements give inaccurate predictions.



Model Evaluation

The model is evaluated using the confusion model evaluation metrics such as accuracy, precision and recall.

The train score:



The test score:



Classification rate of 72% is obtained which is considered as good accuracy for a dataset with 1253 records.

Precision is about being precise and indicating the accuracy of the model implemented. From this it is also understood the correctness of the prediction using the model. Recall indicates that the Logistic Regression model will be able to identify if the candidates would attend an interview or not 93% of the time.

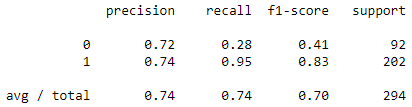
**Implementation using Naïve Bayes**

The Naïve Bayes classifier is based on the Bayes Theorem and this based on conditional probability. The conditional probability helps in calculating the probability that something will happen given that something has already happened.

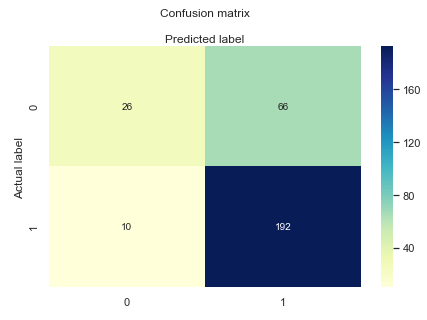
Conditional probability is given by the equation: P(A/B) = P(B/A) \* P(A) / P(B)

The Naïve Bayes classifier assumes all the features are independent to one another. In this project, the model is implemented using the Gaussian Naïve Bayes algorithm. This algorithm is specifically used when the features have continuous values and it is also assumed that all the features are following a Gaussian or normal distribution. For this, the GaussianNB module from the scikit-learn library. The classifier is trained using the training data and uses the fit () method for training it. After doing this, the model is ready to make predictions. As mentioned previously, the predict () method is used to make predictions.

**Classification Metrics**



**Confusion Matrix**



The diagonal elements give accurate predictions and non-diagonal elements give inaccurate predictions.

**Model Evaluation:**

Train Score:



Test Score:



Classification rate of 74% is obtained which is considered as good accuracy for a dataset with 1253 records.

**Chapter 5**

**Conclusion and Recommendation**

The end goal of all data analytics projects is to prefer solutions that are always interpretable and understandable. Interpreting the model is key as it is necessary to verify that the model being built is in line with what is being expected and it allows to create trust with the users. Inorder to improve the accuracy of the model, the project could be further improved by fine-tuning the model to obtain further optimization. It is also important to ensure that the model that is built aligns with the problem that has to be solved. In this project, the problem that needs to be addressed is to predict if a candidate will attend an interview or not. With the implementation of the logistic regression model, the accuracy rate attained is 72.4%. This means that the 72.4%, the candidate will attend the interview. The accuracy rate attained using Naïve Bayes is 74.1% and indicates the prediction percentage of the candidate’s attendance using the model.

The determination of a model that could be considered as ‘good for the project’ depends on the interest of the organization. If the purpose of the model is to provide a highly accurate prediction or decision that is to be used by the business, then a high accuracy should be considered. If interpretation of the business is what is intended, then accuracy score will not be considered but a subjective measure that could help to provide maximum insights would dominate. Since the project aims to predict a candidate’s attendance in an interview, the accuracy of the model could be further improved by using –

1. Add more data

Presence of more data results in better and accurate results

1. Treat missing and outlier values

The presence of missing and outlier values in the training data often reduces the accuracy of the data and would lead to inaccurate predictions.

1. Feature Engineering

Changing the scale of a variable from original scale to a scale between zero and one. This is called data normalization. Grouping values into bins, called as data discretization, is also one way to improve the accuracy score. Creating new features from existing features is called feature creation and helps to unleash the hidden relationship of a dataset.

1. Multiple algorithms

Finding the right machine learning algorithm is the ideal approach to improve the accuracy. The intuition comes with experience and practice.

1. Algorithm tuning

Since machine learning algorithms are driven by parameters. These parameters major influence the outcome of learning process. It is imperative to understand the characteristics of the parameters to tune them so as to improve the accuracy.

1. Ensemble Methods

This is obtained by combining the result of multiple weak models and produce better results. This could be achieved by Bagging (Bootstrap Aggregating) and Boosting. Bagging reduces variance and helps to avoid over fitting. Boosting is a method that is applied to supervised learning by reducing bias and variance.

**References**

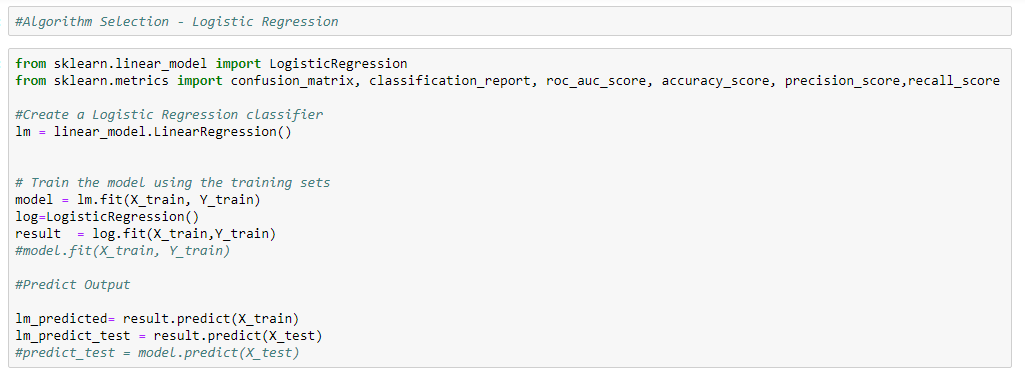
**Appendix A**

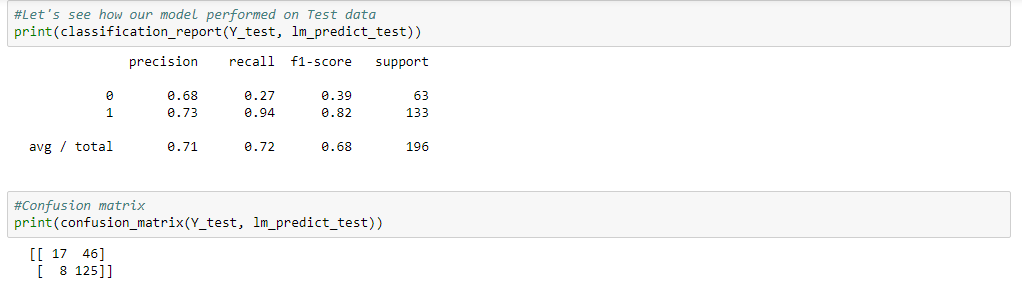
The Candidate Interview Dataset

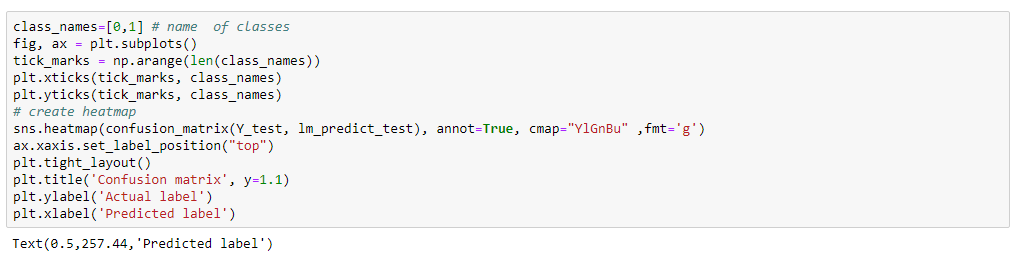
<https://www.kaggle.com/vishnusraghavan/the-interview-attendance-problem>

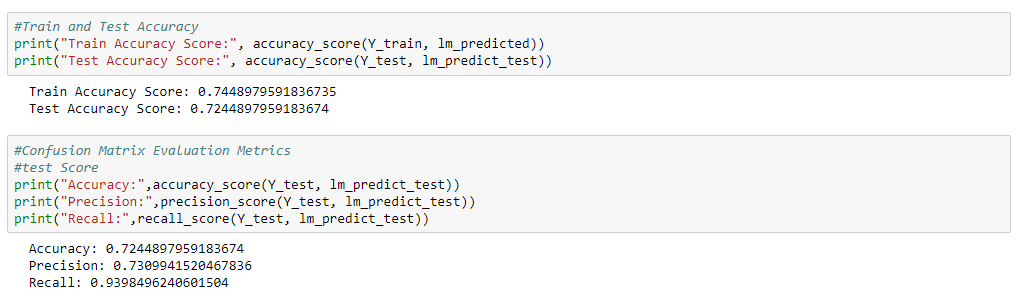
**Appendix B**

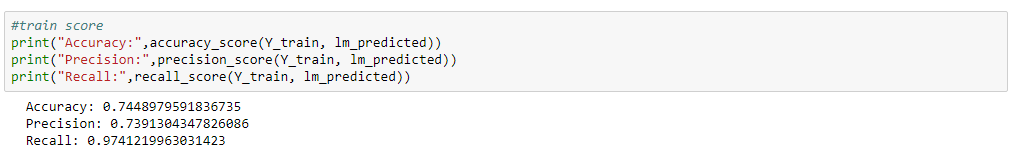
Logistic Regression Model Implementation











**Appendix C**

**Naïve Bayes Model Implementation**



