**TABLE OF CONTENTS**

1. **Project Description 2**
2. **Data Exploration and Pre-processing 3**
   1. **Data Set Exploration 3**
   2. **Data deficiencies 5**
   3. **Data Distribution of Leads and Correlation 5**
   4. **Outliers 6**
   5. **Data Transformation 6**
   6. **Data Partitioning 7**
3. **Model Building 8**
   1. **Logistic Regression 8**
   2. **Support Vector Machines 8**
   3. **K- Nearest Neighbours 8**
   4. **Gradient Boosting 9**
   5. **Decision Tree 9**
   6. **Random Forest 10**
   7. **Neural Networks 10**
   8. **Ensemble Model 11**
      1. **Averaging 11**
      2. **Weighted Averaging 11**
4. **Model Evaluation and AUC Comparision and Conclusion 12**
   1. **Model Evaluation and AUC Comparision 12**
   2. **Conclusion 12**

**1. Project Description**

**“Statement of Academic Honesty:**

The following code represents our own work. We have neither received nor given inappropriate assistance. We have not copied or modified code from any source other than the course webpage or the course textbook. We recognize that any unauthorized assistance or plagiarism will be handled in accordance with Georgia State University's Academic Honesty Policy and the policies of this course. We recognize that our work is based on an assignment created by the Institute for Insight at Georgia State University. Any publishing or posting of source code for this project is strictly prohibited unless you have written consent from the Institute for Insight at Georgia State University.

Online courses are sold by a corporation named X Education, and regular individuals who are interested in their courses visit their website and look through the available courses. The company advertises its courses on numerous websites and search engines like Google, Bing, and others. These potential clients are categorized as leads when they fill out a form with their phone number and email address. Additionally, leads are produced from earlier recommendations. The sales team begins calling, sending emails, and other communications to these leads in an effort to turn them into clients. While some leads are converted, others are not. The conversation rate is not much high, it is near 40%.

Businesses spend a lot of money on their sales and marketing teams. Imagine that X Education receives 100 leads each day but only manages to convert 40% of them. Increasing the conversion rate will increase the process efficiency and, as a result, the company's earnings. The funnel graphic below can be used to depict a typical lead conversion process.

Chart, funnel chart

Description automatically generated

1. **Data Exploration and Pre-processing**
   1. **Data Set Exploration**

The dataset has 9240 rows and 37 columns/features. Out of 37 columns, 4 are float type, 3 of int64 type and 30 are object data type. There are a few nulls, in some columns. We will check that in detail further. Object type columns have descriptive features about the data set. Converted is the target variable and it is of int64 data type. It contains binary values 1/0, which indicates if a prospective customer is converted as a paid customer.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Prospect ID | A unique ID with which the customer is identified. |
| Lead Number | A lead number is assigned to each lead procured. |
| Lead Origin | The origin identifier with which the customer was identified to be a leader. Includes API, Landing Page Submission, etc. |
| Lead Source | The source of the lead. Includes Google, Organic Search, Olark Chat, etc. |
| Do Not Email | An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not. |
| Do Not Call | An indicator variable is selected by the customer wherein they select whether of not they want to be called about the course or not. |
| Converted | The target variable. Indicates whether a lead has been successfully converted or not. |
| TotalVisits | The total number of visits made by the customer on the website. |
| Total Time Spent on Website | The total time spent by the customer on the website. |
| Page Views Per Visit | The average number of pages on the website viewed during the visits. |
| Last Activity | Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc. |
| Country | The country of the customer. |
| Specialization | The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form. |
| How did you hear about X Education | The source from which the customer heard about X Education. |
| What is your current occupation | Indicates whether the customer is a student, unemployed or employed. |
| What matters most to you in choosing this course | An option selected by the customer indicates what is their main motto behind doing this course. |
| Search | Indicating whether the customer had seen the ad in any of the listed items. |
| Magazine | Indicating whether the customer had seen the ad in any of the listed items. |
| Newspaper Article | Indicating whether the customer had seen the ad in any of the listed items. |
| X Education Forums | Indicating whether the customer had seen the ad in any of the listed items. |
| Newspaper | Indicating whether the customer had seen the ad in any of the listed items. |
| Digital Advertisement | Indicating whether the customer had seen the ad in any of the listed items. |
| Through Recommendations | Indicates whether the customer came in through recommendations. |
| Receive More Updates About Our Courses | Indicates whether the customer chose to receive more updates about the courses. |
| Tags | Tags assigned to customers indicating the current status of the lead. |
| Lead Quality | Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead. |
| Update me on Supply Chain Content | Indicates whether the customer wants updates on the Supply Chain Content. |
| Get updates on DM Content | Indicates whether the customer wants updates on the DM Content. |
| Lead Profile | A lead level assigned to each customer based on their profile. |
| City | The city of the customer. |
| Asymmetrique Activity Index | An index and score an assigned to each customer based on their activity and their profile |
| Asymmetrique Profile Index |
| Asymmetrique Activity Score |
| Asymmetrique Profile Score |
| I agree to pay the amount through cheque | Indicates whether the customer has agreed to pay the amount through cheque or not. |
| a free copy of Mastering The Interview | Indicates whether the customer wants a free copy of 'Mastering the Interview' or not. |
| Last Notable Activity | The last notable activity performed by the student. |

From the Data set variables Prospect ID and Lead Numbers are Unique IDs. These will not add any effect on model training. Due to this, we are dropping these columns.

**2.2. Data deficiencies**

Performed Null value treatment on different featuresto identify the percentage of null values in the data set. By dropping those columns which had more than 40% null values, imputing with mean/median where the null values were less. For high null value percentage, we dropped those features because using those features for analysis wouldn’t be of much help because, with the missing data, we cannot infer, but when the null value percentage is less, we imputed those features with mean/median, because it is safe to impute the mean/median for missing as that wouldn’t disturb or bring huge change to the distribution of the data.

We have identified 7 variables having null values of more than 40 %, those are

1. How did you hear about X Education
2. Lead Quality
3. Lead Profile
4. Asymmetrique Activity Index
5. Asymmetrique Profile Index
6. Asymmetrique Activity Score
7. Asymmetrique Profile Score

After removing the columns with Null values greater than 40% then the total count of rows and columns is (9240, 28).

**2.3. Data Distribution of Leads and Correlation**

**Data Distribution of Leads:**

Analyzing the data distribution of the Converted (Target variable). The distribution is not highly skewed but it is nearly balanced. The distribution of not converted to converted is nearly 60 to 40 Percentage.

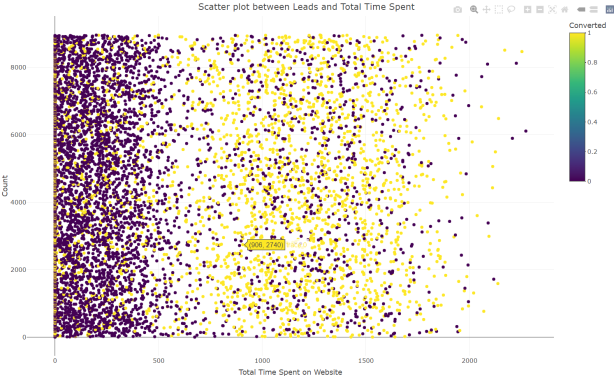
Chart, bar chart

Description automatically generated

**Correlation:**

The data does not contain any duplicate rows; however, we have a column named “Tags” which contains the information about the status of the lead (ex: did not pick up the call, lost to another company,etc). This column cannot be used for modelling purposes since this information can only be available after the sales team made their first call to the customer. So, it won’t be useful to build a model which helps the sales teams to prioritize the potential leads.Correlation coefficients between the numerical variables were calculated to see if there is any variable having a strong linear correlation with the target column.

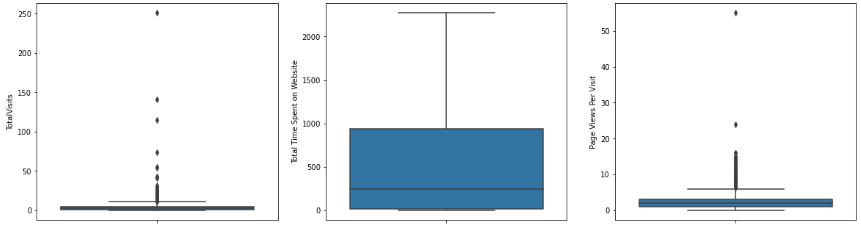
**Graphical user interface, application

Description automatically generated** 

**Correlation Matrix and Plotted Scatter plot for Converted vs Time Spent on Website**

**2.4. Outliers**

We plotted box plots using plotly to identify the outliers for the numerical columns and dropped those values which are below 1% and above 99.00 % Quantile.



The total count of rows after removing outliers is 8953.

**2.5. Data Transformation**

Identifying the distribution of leads respectively to Cities, Specialization, occupation, Lead Source, and Lead Activity.

Chart

Description automatically generatedChart

Description automatically generatedChart, bar chart

Description automatically generatedChart

Description automatically generated

Chart, bar chart

Description automatically generated

1. Several important insights can be gleaned from the plots. Firstly, the majority of the leads are unemployed. Secondly, although the number of employed leads is relatively low, their conversion rate is very high. Finally, leads generated through references exhibit a high conversion rate.
2. Observing the data, it becomes evident that specializations that include "Management" in their title have a greater number of leads, as well as leads that have been converted. Therefore, this variable carries significant value and should not be eliminated. To simplify the analysis, all groups containing "Management" in their title were combined into a single variable.
3. Additionally, some categorical columns exhibit a high degree of skewness, with over 95% of their values being identical. As these variables fail to introduce any meaningful variability to the models, they were excluded from the feature selection process.

Following feature selection, a total of 14 columns were retained for the model-building process. However, as these columns included categorical variables, they needed to be converted into numerical values using encoding techniques. To achieve this, we utilized the fastDummies library's one-hot encoding technique. The resulting process generated 75 columns that can be utilized for the model-building process.

**2.6. Data Partitioning**

The data was split into an 80-20 ratio for training and testing purposes. 80% of the data was allocated for model training, while the remaining 20% was set aside for testing. The sample method was utilized for this purpose.

**3. Model Evaluation**

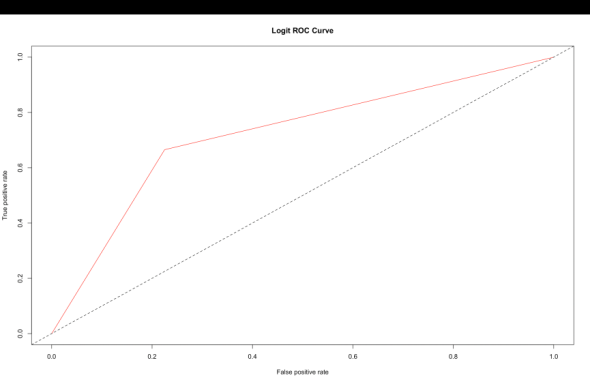
We have performed different models on the data set like Logistic Regression, Support Vector Machines, K- Nearest Neighbour, Gradient Boosting, Decision Tree, Random Forest, Neural Networks, and Ensemble Model.

**3.1. Logistic Regression:**

After applying the Logistic Regression model,



A picture containing table

Description automatically generated 

We have an Accuracy: 73.14% and,

Logistic Regression Area Under Curve(AUC): 0.7201301

**3.2. Support Vector Machines**

After applying the Support Vector Machines Model,



Table

Description automatically generated with low confidence Chart, line chart

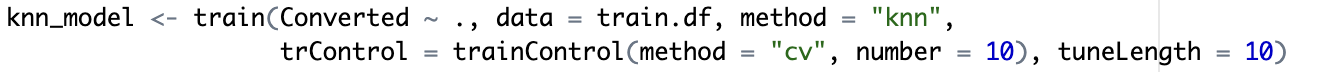
Description automatically generated

We have an Accuracy: 79.12% and,

Support Vector Machines Area Under Curve (AUC) : 0.860056

**3.3. K – Nearest Neighbour:**

After applying the K – Nearest Neighbour,



Table

Description automatically generated Chart, line chart

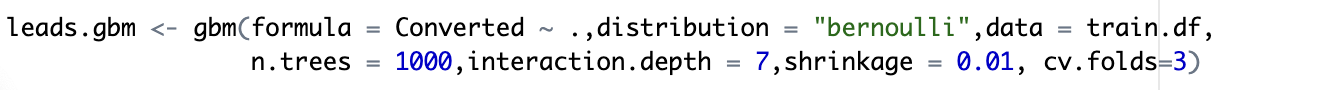
Description automatically generated

We have an Accuracy: 80.07% and,

K- Nearest Neighbour Area Under Curve(AUC): 0.8581028

**3.4. Gradient Boosting:**

After applying the Gradient Boosting,



A picture containing table

Description automatically generated Chart, line chart

Description automatically generated

We have an Accuracy: 83.92% and,

Gradient BoostingArea Under Curve (AUC): 0.9147224

**3.5. Decision Tree:**

After applying the Decision Tree Model,



Table

Description automatically generated with low confidence Chart, line chart

Description automatically generated

We have an Accuracy: 82.69% and,

Decision Tree Area Under Curve (AUC): 0.8543483

**3.6. Random Forest:**

After applying the RandomForest Model, The ‘ntree’ parameter specifies the number of trees to grow in the forest, which in this case is set to 500.



Table

Description automatically generated with medium confidence Chart, line chart

Description automatically generated

We have an Accuracy: 83.53% and,

Decision Tree Area Under Curve (AUC): 0.909002

Table

Description automatically generated

**3.7. Neural Networks:**

After applying the Neural Networks Model, The hidden parameter specifies the number of hidden layers in the neural network, which in this case is set to 3.



Text

Description automatically generated Chart, line chart

Description automatically generated

We have an Accuracy: 82.13% and AUC: 0.8991

**3.8. Ensemble Model:**

This code is building an ensemble model that combines the predictions of multiple types of models. The predictors and outcome variables are first defined, and then three different models are created using the training data. Once these three models are trained, they can be combined into an ensemble model using various methods such as bagging, boosting, or stacking.

Table

Description automatically generated Chart, line chart

Description automatically generated

Ensemble model Accuracy: 79.51% and,

1. k-NN Area Under Curve (AUC): 0.8581028
2. LogisticArea Under Curve (AUC): 0.7201301
3. Decision Tree Area Under Curve (AUC): 0.8543483
4. Gradient Boosting Model Area Under Curve(AUC): 0.9147224

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

1. Averaging ensemble AUC: 0.8912063
2. Weighted averaging ensemble AUC: 0.9059052
3. **Model Evaluation and AUC Comparision**

**4.1 Model Evaluation and AUC Comparision:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **Area Under Curve(AUC)** |
| Logistic Regression | 73.14% | 0.7201301 |
| Support Vector Machines | 79.12% | 0.860056 |
| K – Nearest Neighbour | 80.07% | 0.8581028 |
| Gradient Boosting | 83.92% | 0.9147224 |
| Decision Tree | 82.69% | 0.8543483 |
| Random Forest | 83.53% | 0.909002 |
| Neural Networks | 82.13% | 0.8991 |
| Ensemble Model | 79.51% | 0.8912063 |

From the above Table, we see that Gradient Boosting Model had given the Highest Accuracy of 83.92% and the Highest AUC value of 0.9147224

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.
2. Last Notable.Activity\_SMS.sent
3. Lead.Origin\_Lead.Add.Form
4. Pages.Views.Per.Visit
5. TotalVisits
6. What.is.your.current.occupation\_Working.Professional
7. When the lead source was:
8. Google
9. Direct traffic
10. Organic search

**4.2 Conclusion**

1. Using these key takeaways, the sales team can increase their sales, without making modifications to their already existing process. As this will give the company the scope to identify prospective customers to prioritize and they can operate within their current capacity without needing any additional resources.
2. And in future, the new data generated can be fed to the model for it to learn better and give better predictions about the features, as the feature importance might keep changing according to time and the sales team can keep up with the changes.