Lead Scoring Case Study

Steps Followed:

- Data Reading and Understanding
- Data Cleaning
- Data Visualization
- Dummy Variables creation
- Test Train data Split
- Scaling Numerical Variables
- Model Building
- Evaluating model on Train dataset
- Calculating efficiency metrics and plotting Roc Curve
- Finding optimal cut off using sensitivity, specificity and accuracy
- Calculating Precision and Recall
- Evaluating the model on test set

Data Sourcing

- Loaded Data from the leads.csv into data frame using pandas library.

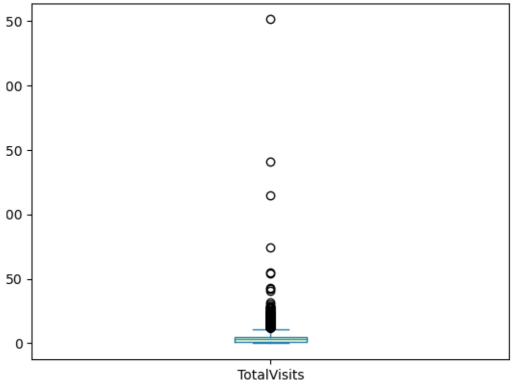
Data Cleaning

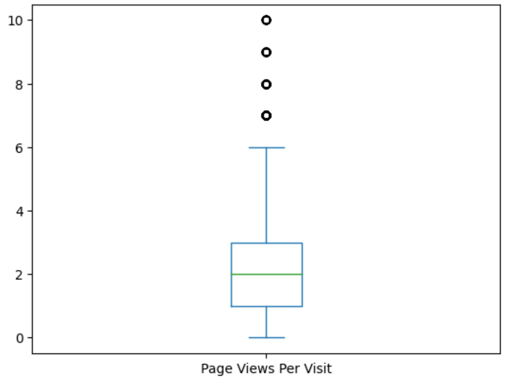
- Null Values in the columns are handled following below steps:

1. Dropping the column if percentage is more
2. Dropping the rows.
3. Imputing them with mode value.

Data Visualization

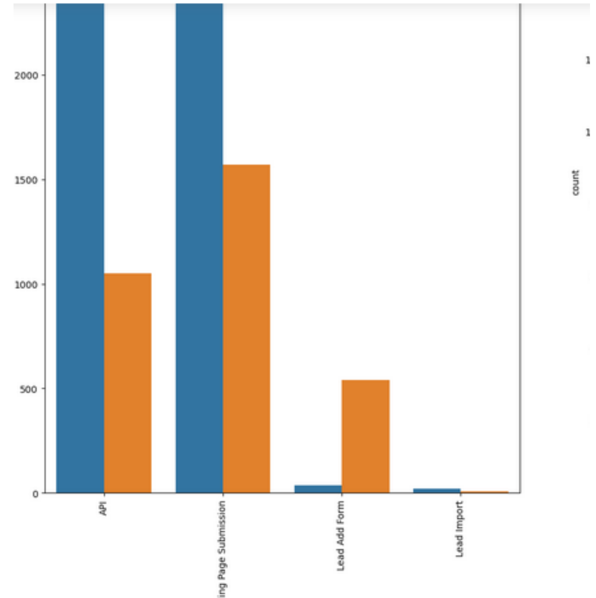
- Outliers are handled using univariate analysis and values above 95% quantile are removed for 2 numerical variables.





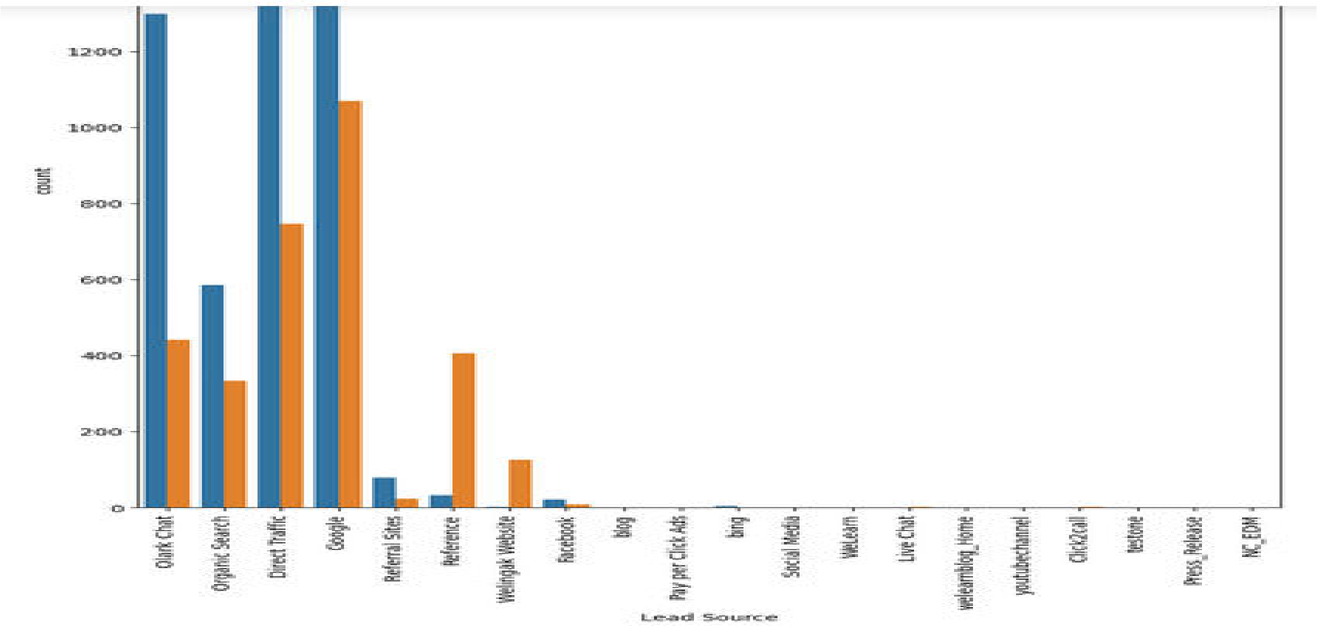
Bivariate Analysis

- Leads originated from lead add form are less ,but conversion rate is more, so need to communicate with the people who is on add form to convert them.
- Landing Page submission ,API are also having more leads. But need to focus on increasing conversion rate of this type.



Bivariate Analysis

- Lead conversion rate for the leads whose Lead Source is Reference ,Welingak website is high comparatively, so need to focus on getting more leads from this source.

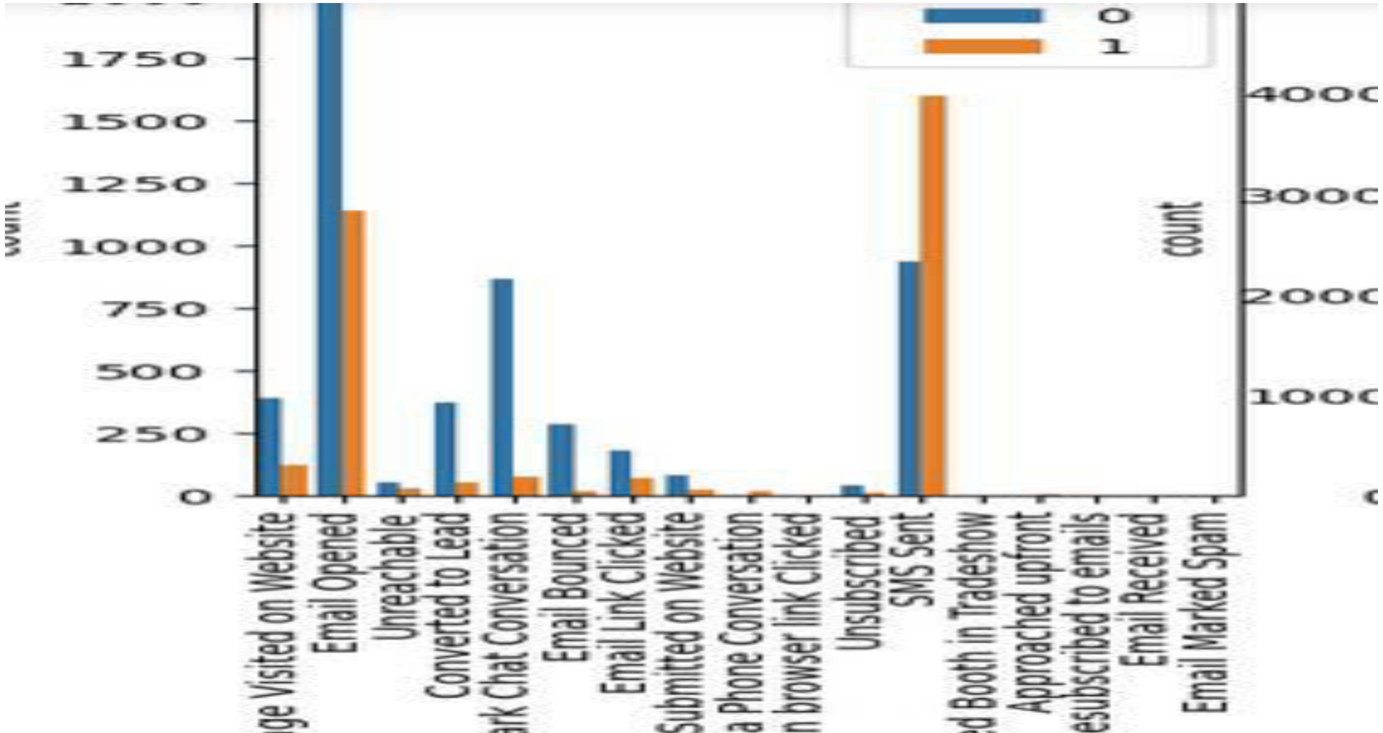


Bivariate Analysis

- Last Activity:

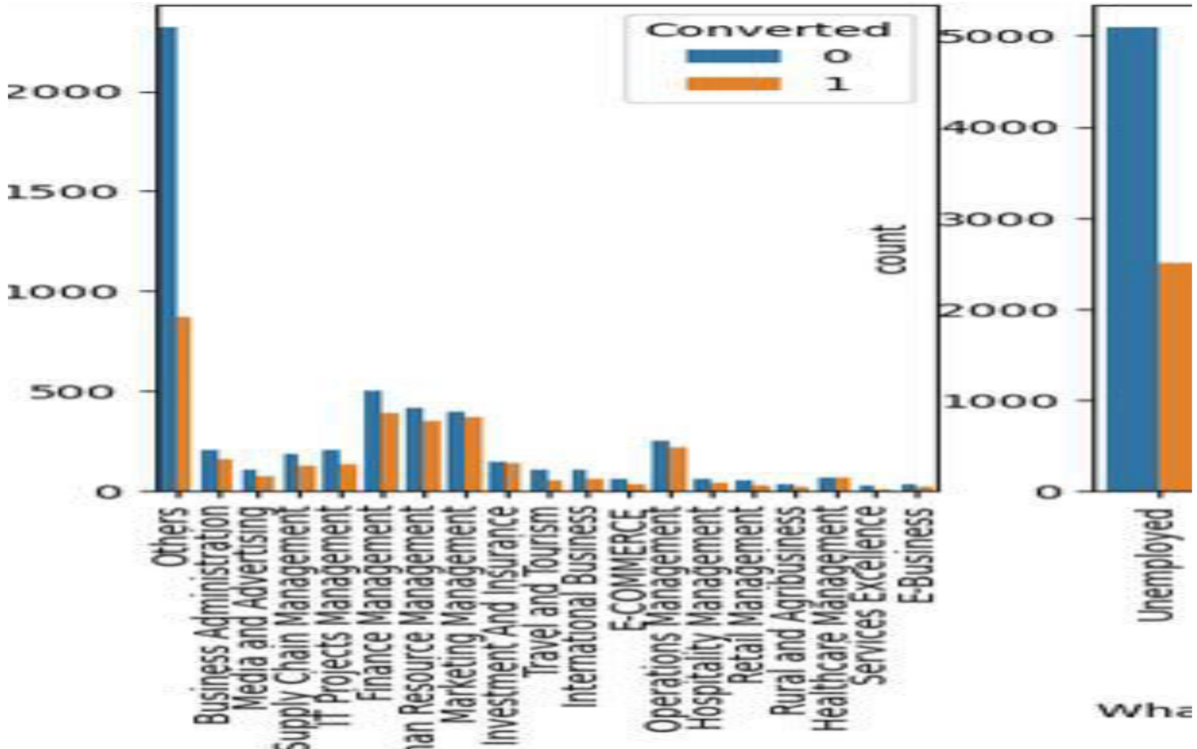
SMS Sent has high conversion rate, need to focus on getting more leads.

Email Opened has high leads but conversion rate is less, need to focus on converting them.



Bivariate Analysis

- **Specialization**:
- conversion rate of leads from finance management, Human resource management and marketing management ,operations management is high, so need to focus on getting more leads from these specializations.

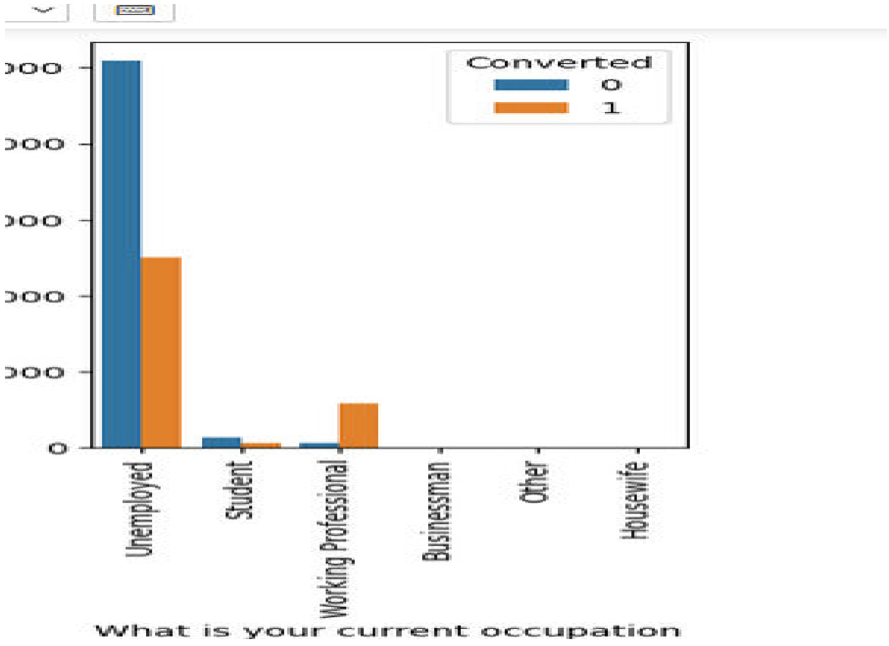


Bivariate Analysis

- What is your current occupation:

Working professional has high conversion rate : so need to focus on getting more leads of this category

Origination rate of unemployed people is more. Need to communicate on increasing their conversion rate.

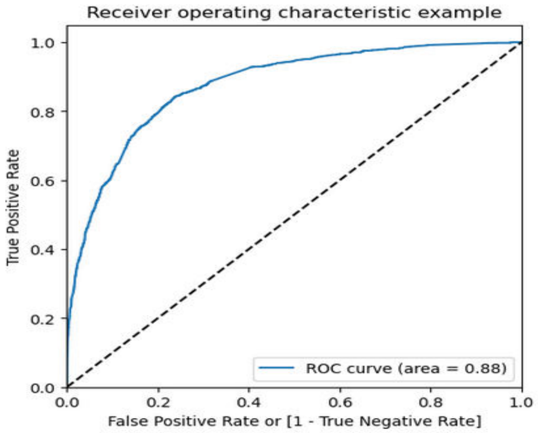


Dummy Variables Creation:

- Dummy Variables are created for the categorical variables Lead Source, Do Not Email, Lead Origin,
- Last Activity, Specialization, What is your current occupation, City, Last Notable Activity,

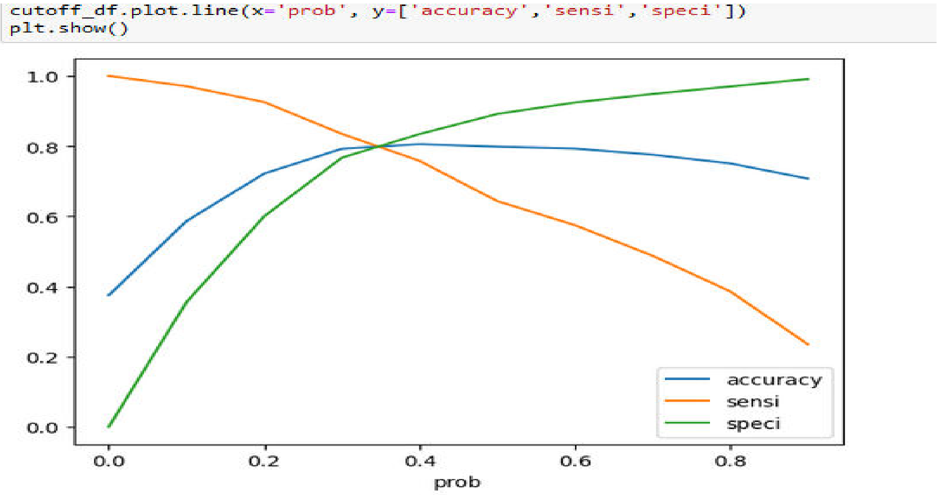
Model building

- Data is split into test and train data set.
- Train data set numerical variables are scaled to feed into model.
- Model is built using RFE. Arrived to a final model basing on p-value and VIF values.
- Model is evaluated on train data set and target variable probabilities are calculated.
- Arbitrary value of 0.5 is chosen and target variable is predicted.
- ROC Curve is plotted.



Finding Optimal Cut off

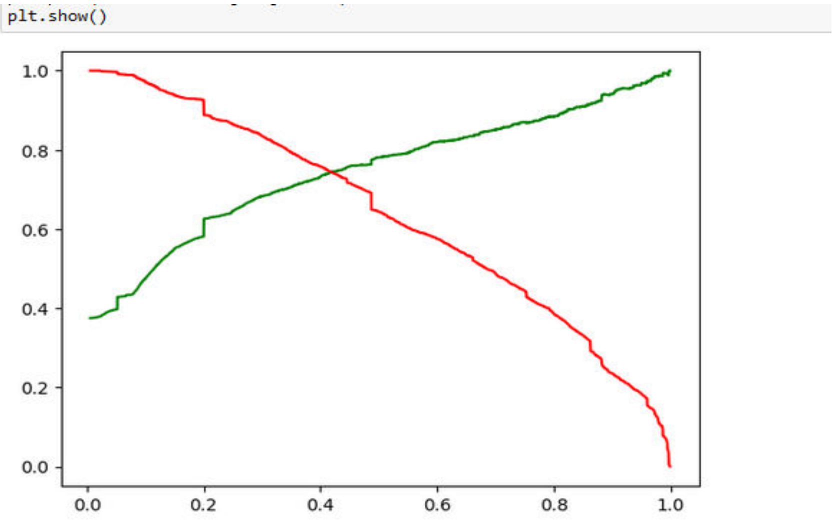
- Optimal cut off which is 0.36 is found using sensitivity, specificity and accuracy curves plot against probabilities
- All the efficiency metrics are once again calculated with optimal cutoff.



```
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```

Precision and Recall

- Precision and Recall are calculated and plotted.



```
plt.show()
```

Evaluating model on Test Set

- Model is Evaluated on test set
- Accuracy of model is 82%
- Sensitivity is 81%
- Specificity is 82%

Inferences from Model

- Company should focus on making calls for the leads :
- Whose Lead Source is Welingak Website/Olark Chat.
- Whose Lead Origin is Lead Add Form
- Whose Last Activity is either a Phone Conversation/SMS Sent.
- Who spent more time on the website.
- Who are working professionals.