# **Apptio Assessment (Data Scientist): LEAD SCORING CASE STUDY[¶](http://localhost:8888/notebooks/Desktop/%23 Apptio Assessment (Data Scientist): LEAD SCORING CASE STUDY.ipynb" \l "Apptio-Assessment-(Data-Scientist):-LEAD-SCORING-CASE-STUDY)**

Submitted By: Sanidhya Singh

1. How long did it take you to solve the problem?

Ans) Around 6 hours

1. What software language and libraries did you use to solve the problem?

Ans) Language Used: Python

Packages used:

* numpy
* pandas
* Matplotlib
* Seaborn
* Sklearn
* Statsmodels
* Warnings
* sklearn.model\_selection
* sklearn.feature\_selection
* sklearn.linear\_model
* statsmodels.stats.outliers\_influence
* sklearn.metrics

1. What steps did you take to prepare the data for the project? Was any cleaning necessary?

Ans)

* Remove the columns which have unique values
* Remove the columns which didn’t show any variation
* Remove the columns with more than 70% missing values
* Changed “Select” labels to null values.
* For remaining missing values, we imputed values with maximum number of occurrences for a column.
* Changed the multicategory labels into dummy variables and binary variables into ‘0’ and ‘1’.
* Outlier handling.
* Removed all the redundant and repeated columns

1. What algorithmic method did you apply? Why? What other methods did you consider?

Ans) Logistic Regression. It was a binary problem, and logistic regression works well with binary problems. We created our model with rfe count 19 and 15 and compared the model evaluation score like AUC and choose our final model with rfe 19 variables as has more stability and accuracy than the other.

1. What features did you use? Why?

Ans)

Total Time Spent on Website

Lead Origin\_Lead Add Form

Lead Source\_Olark Chat

Lead Source\_Welingak Website

Last Activity\_Email Bounced

What is your current occupation\_Working Professional

Tags\_Busy

Tags\_Closed by Horizzon

Tags\_Lost to EINS

Tags\_Ringing

Tags\_Will revert after reading the email

Tags\_switched off

Lead Quality\_Not Sure

Lead Quality\_Worst

Last Notable Activity\_Modified

Last Notable Activity\_Olark Chat Conversation

Last Notable Activity\_SMS Sent

1. How did you train your model? During training, what issues concerned you?

Ans)   
Data Preparation:

1. Split the dataset into train and test dataset and scaled the dataset.
2. b. After this, we plot a heatmap to check the correlations among the variables.
3. Found some correlations and they were dropped.
4. 4. Model Building:
5. We created our model with rfe count 19 and 15 and compared the model evaluation score like AUC and choose our final model with rfe 19 variables as has more stability and accuracy than the other.
6. For our final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity.
7. We found one convergent points and we chose that point for cutoff and predicted our final outcomes.
8. We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs.
9. Prediction made now in test set and predicted value was recoded.
10. We did model evaluation on the test set like checking the accuracy,

recall/sensitivity to find how the model is

1. We found the score of accuracy and sensitivity from our final test model is in acceptable range.
2. We have given lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads.
3. How did you assess the accuracy of your predictions? Why did you choose that method? Would you consider any alternative approaches for assessing accuracy?

Ans)

Curently Used: Confusion Matrix, AUC.

Alternative: Z Score

1. Which features had the greatest impact? How did you identify these to be most significant?

Ans)

* Last Notable Activity\_Had a Phone Conversation
* Lead Origin\_Lead Add Form
* What is your current occupation\_Working

Professional

Identified with the help of Rfe ranking, p score, z score.