LEAD SCORING CASE STUDY

SUBMISSION

Group Name: United By Chance

1. Ashish Gaurav
2. Ayushman Priye
3. Chandan Agrawal
4. Rahul SP

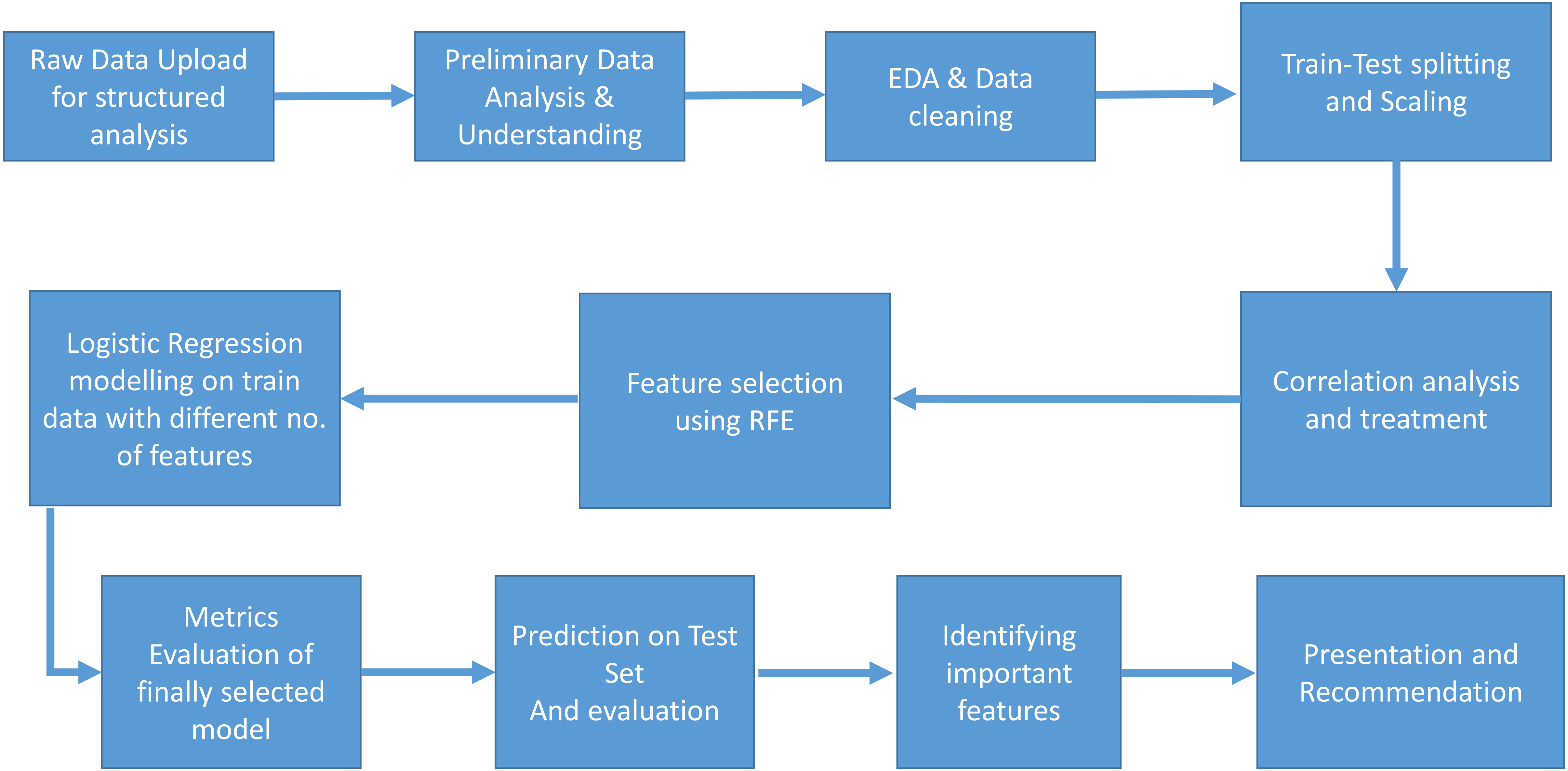
<Abstract - Objective and Problem Statement>

- Our company, X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- As the lead conversion rate is very poor, to make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone..
- We thus need to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- In other words, we need to understand which factors are important in increasing the probability of a lead getting converted and based on these factors build a model which assigns a probability score to each prospect. Higher this score, higher the probability of that prospect getting converted into a client.

<Some Preliminary Observations/Notes>

- The dataset contains 9240 records with 37 columns containing detailed information about the loans applications.
- The column ‘Converted’ is the target variable which tells whether the lead converted into customer or not.
- A large number of columns are ‘Categorical’ in nature thus a large no. of dummy variables need to be created.
- Many columns have single value for all the records (like 'Get updates on DM Content’). Such columns can be dropped as they do not add any value.
- In many columns, ‘Select’ appears as a value which indicates that the lead didn’t selected any available option for that field. Based on the column, either it has been treated as a Null value or handled through Dummy method.
- Some of the variables are highly coreeleated and thus can be treated accordingly.

<Approach followed>



Raw Data Upload for structured analysis

Preliminary Data Analysis & Understanding

EDA & Data cleaning

Train-Test splitting and Scaling

Correlation analysis and treatment

Feature selection using RFE

Logistic Regression modelling on train data with different no. of features

Metrics Evaluation of finally selected model

Prediction on Test Set
And evaluation

Identifying important features

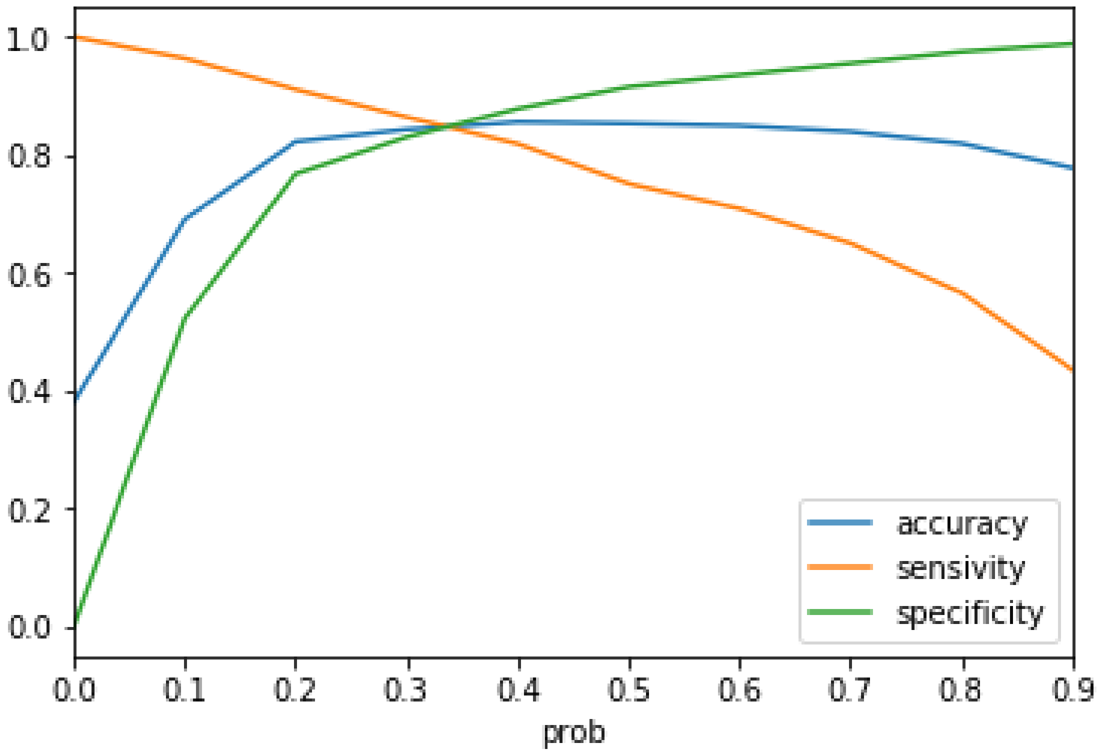
Presentation and Recommendation

<Results>

- Though we have about 34 independent variables to predict a lead will be converted or not, a lot many variables are not that useful and do not influence the conversion probability much.
- The top three variables whic contribute most towards the probability of a lead getting converted are:
  - Lead Source
  - Lead Quality
  - Last Notable Activity
- Besides these 3 some other important variables are:
  - Country
  - Current occupation
  - Total Time Spent on Website
  - Assymetric Activity Index

<Results>

- The ‘Accuracy vs Sensitivity vs Specificity’ plot for the model is shown below :



So, the optimal cut-off comes out to be about 0.35, but it can be changed as per the situation.

<Recommendations>

- As mentioned, the top three variables whic contribute most towards the probability of a lead getting converted are:
  - Lead Source
  - Lead Quality
  - Last Notable Activity

- The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are:
  - LeadSource_Welingak Website – it means if the lead source is ‘Welingak Website’, there is higher probability of lead conversion
  - LeadQuality_High in Relevance – it means leads with ‘High in Relevance’ in Lead Quality have higher probability of conversion
  - LeadQuality_Low in Relevance - it means leads with ‘Low in Relevance’ in Lead Quality have higher probability of conversion

- So, if the lead is sourced from ‘Welingak Website’ or the quality of lead is tagged as ‘High in Relevance’ or ‘Low in Relevance’, then such lead should be given more focus and should be contacted for sure.

<Recommendations>

- If we can afford to call a large number of leads, it will be helpful if the model doesn’t predict a potential lead as a non-potential one. So, the cases of ‘False Negatives’ should be as minimum as possible. In other words, the Sensitivity should be high. So, we can decrease the cut off probability to predict whether the lead will be Converted or not. By doing so, it is less likely that we miss a true potential lead. As we can afford to connect with large number of leads, even if the call goes to leads with very low probability of conversion, it is fine.
- If we don’t want to make calls unless it’s extremely necessary (i.e unless there is very high chance of the lead getting converted) , it will be helpful if the model doesn’t predict a non-potential lead as a potential one. So, the cases of ‘False Positives’ should be as minimum as possible. In other words, the Specificity should be high or False Positive Rate should be low. So, we can increase the cut off probability to predict whether the lead will be Converted or not. By doing so, it is less likely to give an unnecessary call (i.e. calling a lead which has low probability of getting converted). Though in this case some true potential leads will also be missed but given the situation we can afford to do so.