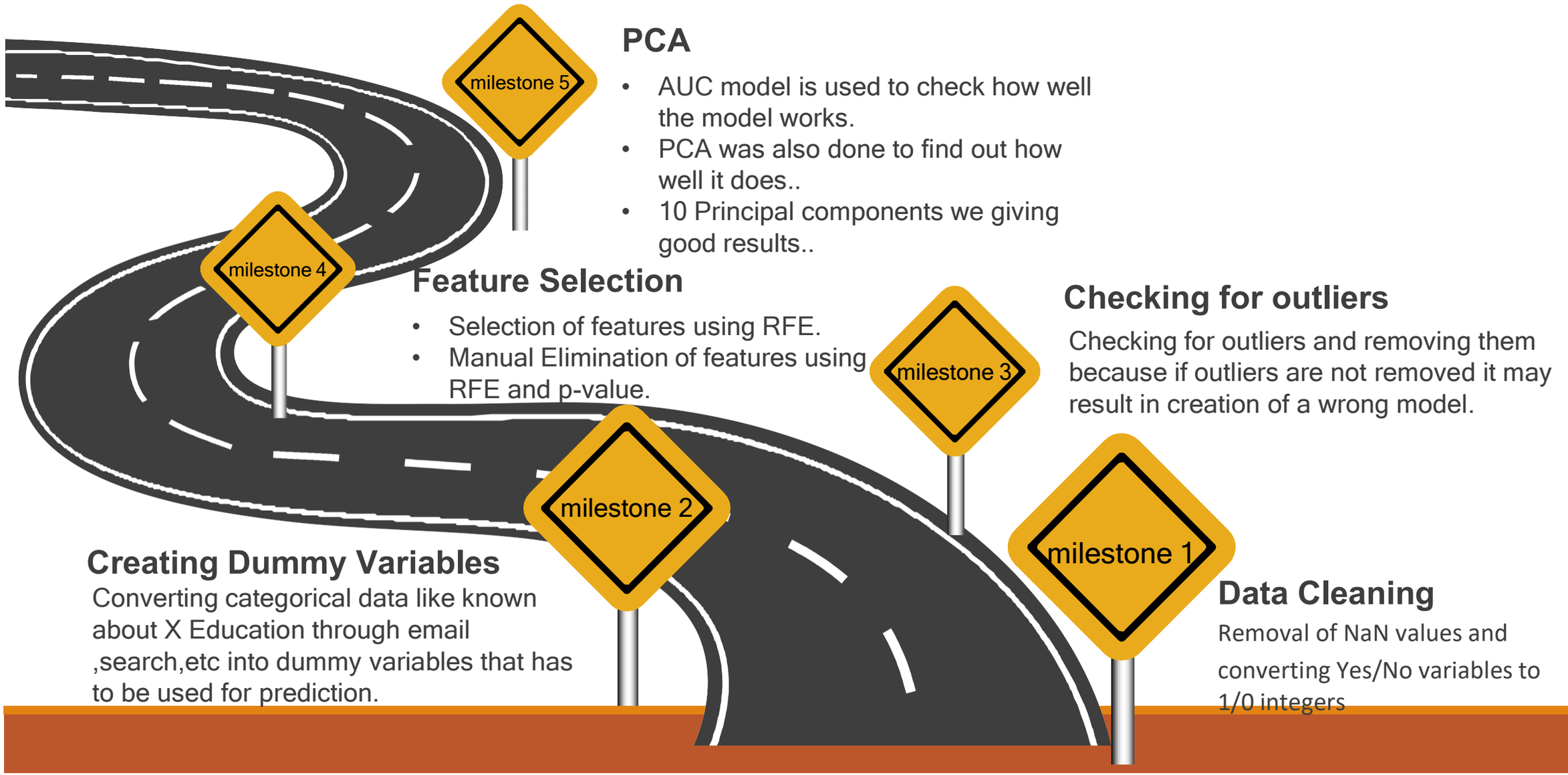
LEAD SCORING OF AN EDUCATION COMPANY

Introduction

- The company X-Education has got a data set containing the information about a set of people who have given them the information through various sources and also interacted with their website.
- These are called leads.
- If we could identify the leads that are more likely or less likely to convert into a paying customer then the company could spend its resources wisely.
- In order to do this we made a logistic regression model which outputs the probability of a lead becoming a paying customer given the lead's information as input.
- The steps involved in modelling are briefly explained in the next slide.

Roadmap to the construction of Regression Model



PCA

- AUC model is used to check how well the model works.
- PCA was also done to find out how well it does..
- 10 Principal components we giving good results..

Feature Selection

- Selection of features using RFE.
- Manual Elimination of features using RFE and p-value.

Checking for outliers

Checking for outliers and removing them because if outliers are not removed it may result in creation of a wrong model.

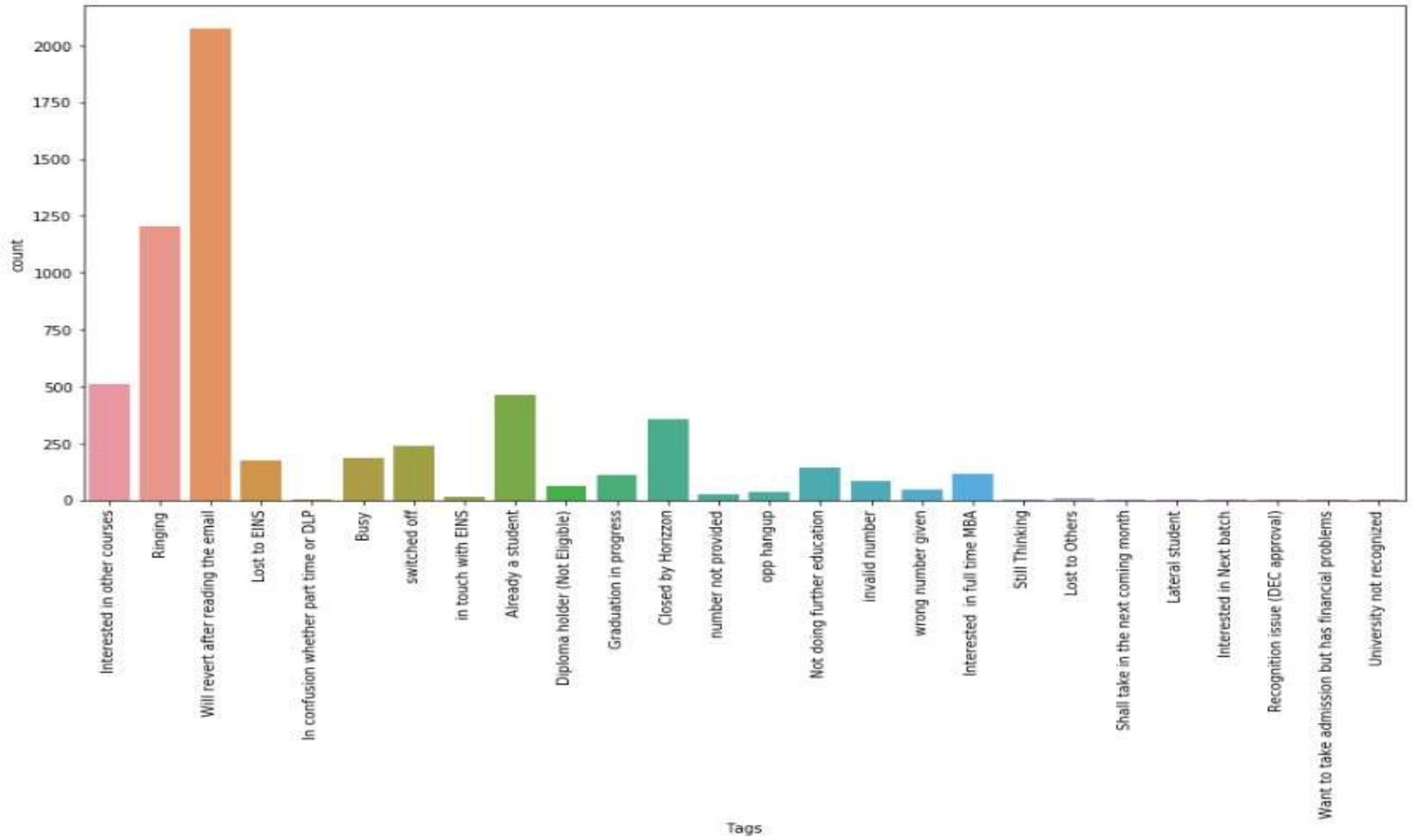
Creating Dummy Variables

Converting categorical data like known about X Education through email ,search,etc into dummy variables that has to be used for prediction.

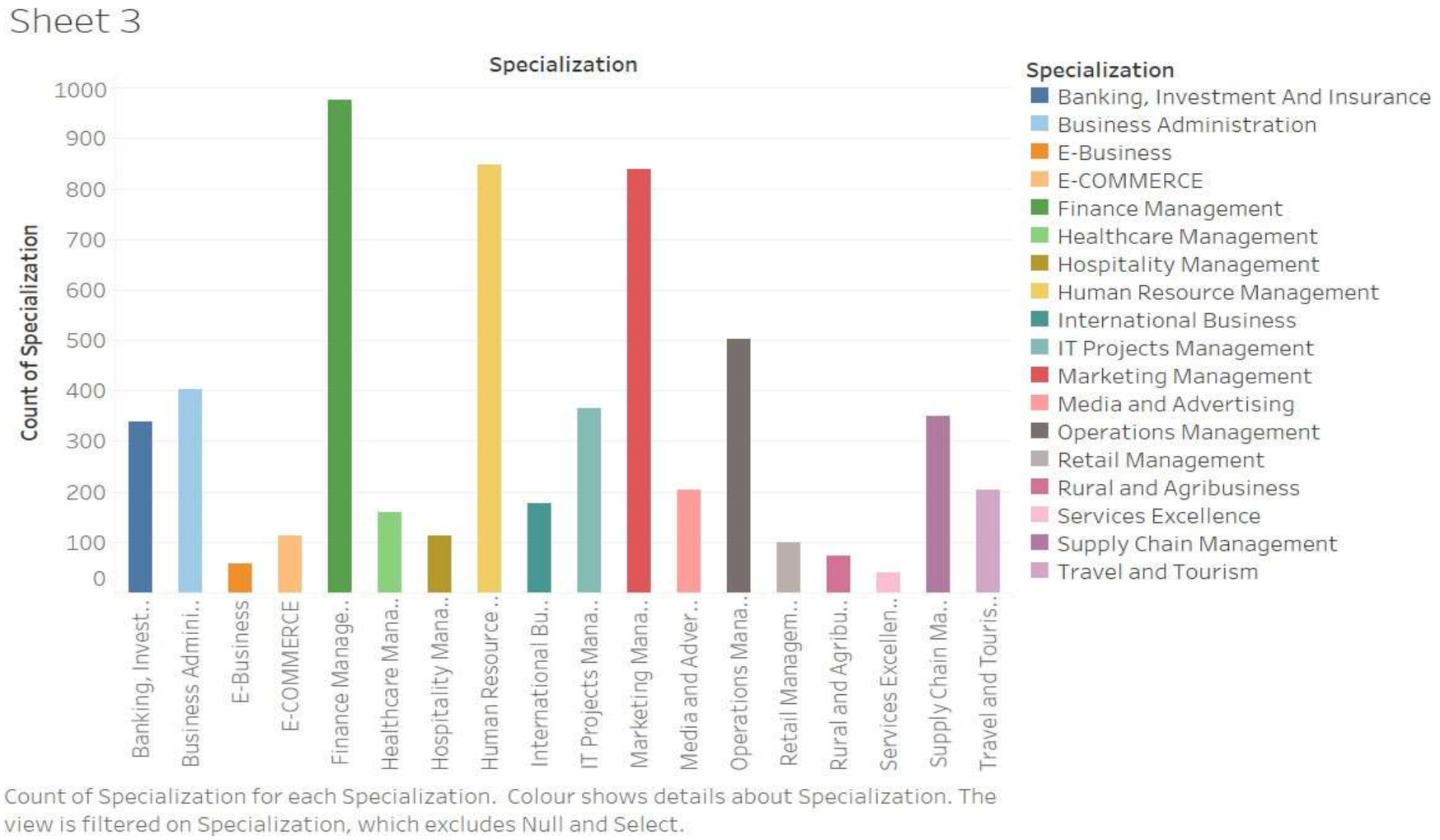
Data Cleaning

Removal of NaN values and converting Yes/No variables to 1/0 integers

The most common tag is “will revert back”.
It maybe the case that lead has not entered any specialization if his/her option is not available on the list.



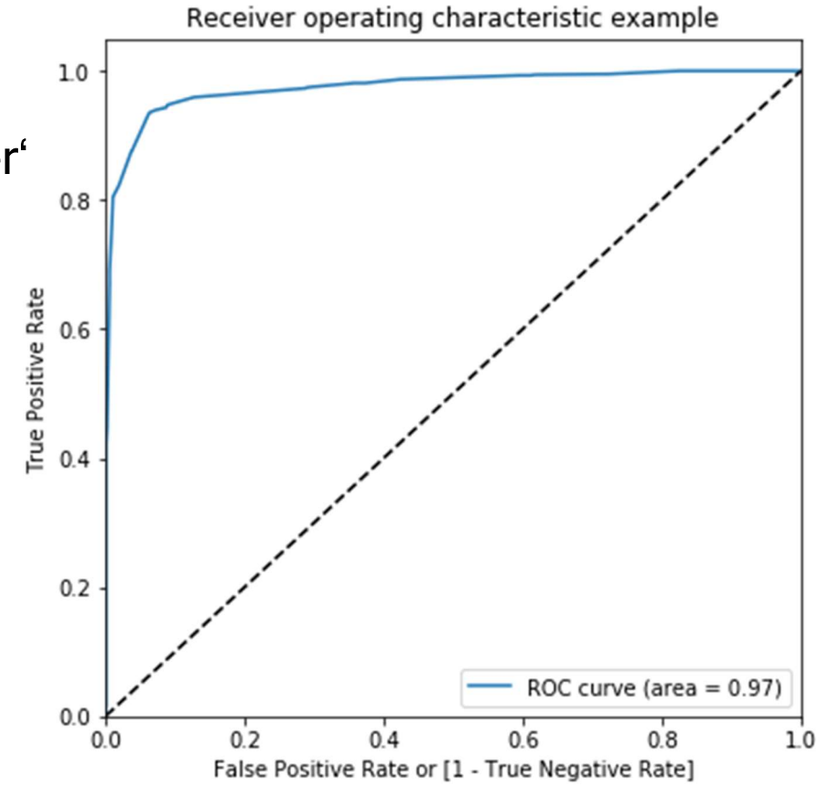
The course most people enrol in is Financial management



Count of Specialization for each Specialization. Colour shows details about Specialization. The view is filtered on Specialization, which excludes Null and Select.

Important variables for model building

1. 'Lead Source_Welingak Website‘
2. 'Last Activity_SMS Sent‘
3. 'What matters most to you in choosing a course_Other‘
4. 'Tags_Busy‘
5. 'Tags_Closed by Horizzon‘
6. 'Tags_Lost to EINS‘
7. 'Tags_Ringing‘
8. 'Tags_Will revert after reading the email‘
9. 'Tags_number not provided‘
10. 'Tags_switched off‘
11. 'Lead Quality_Worst‘
12. 'Last Notable Activity_Modified‘
13. 'Last Notable Activity_Olark Chat Conversation'

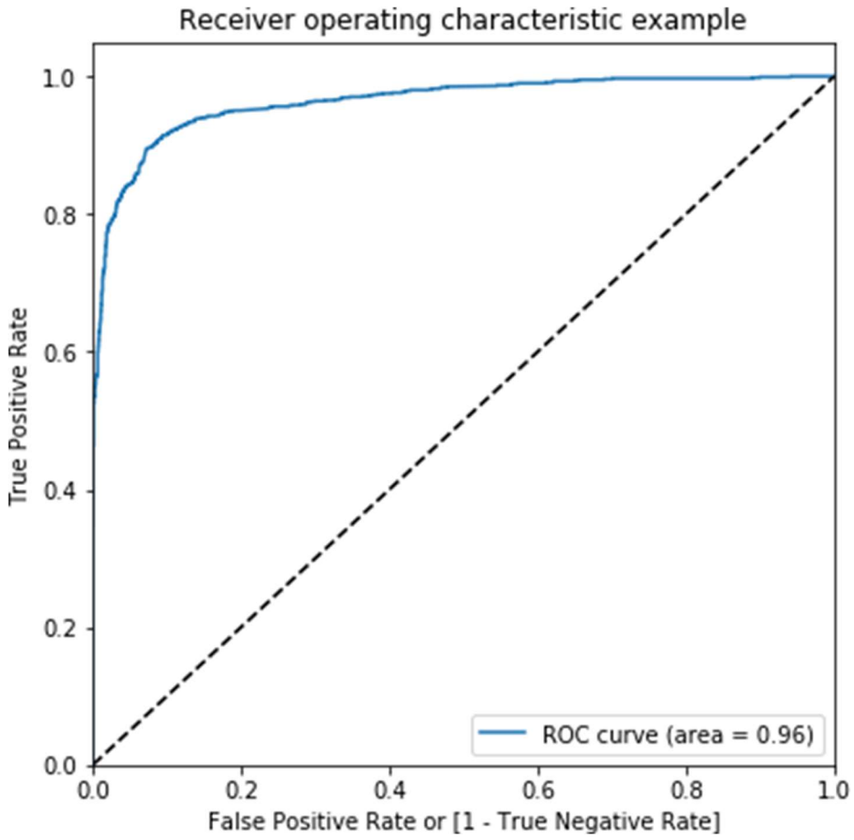


ROC curve for the model build

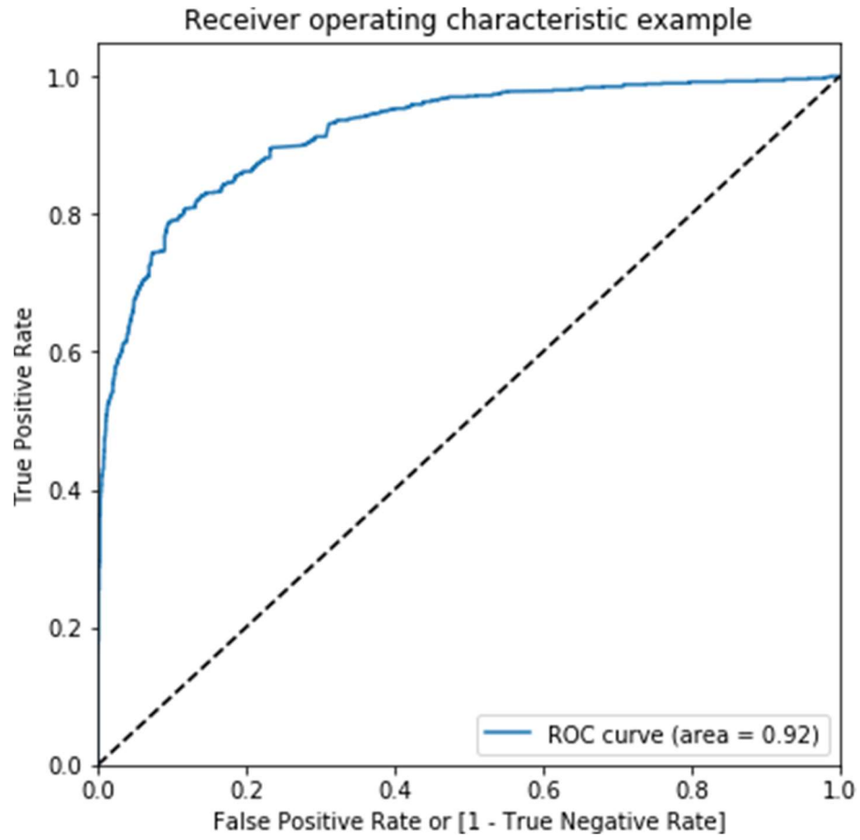
Important Results

- The model that was built using RFE and manual feature elimination had the area under ROC curve(AUC) as 0.97 which is very close to 1.
- Hence that model gave very good results.
- But we also built a model by selecting features based on PCA to see how well it does and how many principal components could explain good variance.
- It had the AUC of 0.96.
- With just 10 features it gave a AUC of 0.92

ROC curves - PCA



ROC with 95% variance



ROC with 10 components

Problem 1

Problem Statement

Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

Solution

Top three variables which contribute most towards the probability of a lead getting converted are:-

1. Tags_Lost to EINS
2. Tags_Closed by Horizon
3. Tags_Will revert after reading the email

Problem 2

Problem Statement

What are 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?

Solution

Top three categorical/dummy variables which must be focused most on in order to increase the probability of a lead getting converted are:-

1. Tags_Lost to EINS
2. Tags_Closed by Horizon
3. Tags_Will revert after reading the email

Problem 3

Problem Statement

X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

Solution

When interns are available the company has more resources to spend on nurturing the leads hence we can choose a lower threshold value on the probability of the predicted lead conversion value. This implies that even though the probability of lead conversion is low we try our best to convert by making phone calls and pursue them using the labor force. This mean that we are not efficient but since the company wants to be aggressive this works well.

Problem 4

Problem Statement

Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Solution

In this case we have less resources available for nurturing and need to be very highly efficient. So the company should concentrate on the leads that have a very high chance of becoming a paying customer and hence we choose a high threshold value on the probability outcome of the logistic regression model that we have built. Now we only concentrate on the leads that have higher probability of conversion.

Take Away

1. The company should focus only on the higher probability leads when the resources are low and go for lower probability leads too when they have good resources.

2. We see that a manual features selection is a lengthy and time consuming process, similar results can be obtained using PCA

3 The ROC curve for a model with manual feature selection and 95% percent variance produce similar results, hence we can say that PCA gives similar results with less effort

Thank You