Summary Report

Problem:

Company X-education gets leads from various sources but has poor lead conversion. They want identify ‘hot leads’ (which are most likely to convert) so that the sales team can focus on them and not waste time and other resources on leads which are less likely to convert

Objective:

To build a logistic regression model which assigns a score to every lead. This lead score will range from 0-100, where increasing score indicates increasing likelihood of conversion. Based on a cutoff of our requirement we can then predict which leads will convert. The sales team will then approach these leads and move forward. There are 2 terms which become important here -

1. Precision - Out of all the leads which were predicted to convert, how many did convert.
2. Recall - Out of all the leads that converted, how many could we predict would convert.

Model Building Summary:

We went through the general template for model Gebuilding - Data cleaning, Exploratory Data Analysis, Data Preparation, Model Building, Prediction, Evaluation, Optimization, Interpretation and Conclusion. However, there are certain specific points we kept in mind while building a model for this problem that X-education faces -

1. Data generated by the sales team after contacting the lead had to be removed from our dataset. This is because our model is to be deployed on leads to identify the “hot leads” before the sales team contacts them. This allows them to focus on those customers which are likely to convert.
2. Since a lead score had to be created which was to convey the probability that a lead is likely to convert, the model should only give out the probabilities of leads converting, and not the final prediction if the lead will convert or not.
3. The final prediction on if the lead will convert or not was to remain in our hands, i.e. there should be removed for manual input. We decide the cutoff based on how much precision or recall we want.

Result:

1. Leads brought in through references, other smaller websites (mostly Wellingak) are significantly more likely to convert. Olark chat and google follow behind (in this order)
2. If a lead chooses to not receive emails, this perhaps shows that they are not as serious about the course and it is more likely that they will not convert.

3. If a lead is a working professional, it is significantly more likely that they will convert as opposed to someone who isn’t working (unemployed, students and others).
4. Customers who spend more time on our website show significantly greater chances of being 'hot leads'
5. Leads which visit our site more, are more likely to buy our course (perhaps signalling growing interest and conviction).
6. Leads which view more pages per visit on our site are less likely to convert. Perhaps due to paralysis by analysis

Recommendations:

1. Get more traffic to Wellingak and establish a better network of references
2. SEO on google, domain specific community participation on facebook, LinkedIn etc.
3. Tailor course to working professionals.
4. Improve website UI.