Lead Scoring
Final Report

Project Overview
In this project, I created a model that uses sales data to assign a numerical score to incoming leads, indicating the probability of conversion for each lead. The data contained several categorical and numerical features to describe lead characteristics, behaviors, and perceived quality. The best performance was achieved with an XGBoost model, optimizing log loss to emphasize the importance of accurate predicted probabilities. The area under the ROC curve for the final model was 0.94, with recall for converted leads at 50% probability threshold of 0.81.

Using this model, the company can increase their lead conversion rate by adapting sales strategies to an individual lead's score. The model also identified the most important features predicting conversion, which the company can use to update their marketing strategies to focus on attracting leads that are more likely to convert.

Problem Statement
In nearly any client-focused business, the acquisition of new customers is a high priority requiring a large investment of time, money, and labor. In order to have the best outcomes, new leads generated (say, by a marketing team) must be appropriately nurtured (by sales staff) until the sale is made, and we can say the lead has been converted. Naturally, however, not all leads are of the same quality, and spending valuable time on hopeless leads can be an inefficient use of energy and resources.

In my personal sales experience at a private adult education provider, I've wasted plenty of time writing multiple emails and calls to potential students that would go unanswered. As the only sales advisor in our small branch, that was time better spent drumming up new leads or ensuring a strong relationship with current students that I had already converted.

One possible solution to this problem is a reliable model of lead scoring. Incoming leads can be assessed based on multiple features, including data explicitly collected from the potential client (such as personal identifiers like job or location, or answers to questions about their specific motivation in choosing a product) and data implicitly gathered from them (such as time they've spent on the website, or whether they have responded to communication). Each lead is then given a score indicating how likely they are to convert to a sale based on the available information.

With robust lead scoring, a company can simultaneously pursue two pathways to better sales outcomes:

1. **Better Lead Conversion Rate**
   Attention to lead score will make sure sales staff are prioritizing the leads with good change of conversion, and conversely deprioritizing unlikely sales. This should lead to a greater overall conversion rate, as good leads aren't likely to be accidentally ignored.

2. **Better Lead Acquisition**
   By identifying the features most important to determining lead score, the model gives evidence for prioritizing certain markets over others. For example, if school applicants from major cities have a higher conversion rate than rural ones, marketing efforts should be focused on those areas.

Using the labelled lead data we've collected, we can design a model that will identify the most predictive features and assign a probability of conversion for each new lead, which we can then convert into a score from one to ten. Armed with this score, company sales management can then make an informed decision of how to assign and treat incoming leads. Meanwhile, by looking at the percentage of high- and low-scoring leads, the company can adjust marketing techniques to try to maximize the number of strong leads generated.

Data Wrangling

As real data on leads is not generally available for free use (largely to protect client privacy), I used data simulated for educational purposes by UpGrad IIIT-B, accessed via kaggle.com. The data represents lead information for X Education, an online adult education company.

The data contained 37 columns, with records for 9,240 unique leads. Although the Kaggle data source indicated in the description that "The typical lead conversion rate at X Education is around 30%," the actual conversion rate in the dataset was 38.5%. This may indicate that either the sample is not random (conversions were upsampled as the target population), or that the actual conversion rate varies considerably. As a result, the company must closely monitor model performance once applied to new data, to make sure the results are what is expected.

On inspection, some columns included in the original dataset were found to be unhelpful in distinguishing leads, and were therefore dropped. The columns 'no_call', 'news_article', 'forums', 'newspaper', and 'digital_ad', all of which contained Boolean markers of lead preferences, had over 99.9% uniformity.

The majority of the other columns contained categorical values. Some columns had a high percentage of null values. However, I decided it would be best to keep this information for our model just in case the lack of data for a particular lead in fact indicates something meaningful about their probability of conversion. For instance, if a lead declines to provide their current occupation might not be very serious about their application. So, I recoded null values in all columns as a separate 'Unknown' category.

One column in particular was difficult to interpret, however. The 'tags' column contained "Tags assigned to customers indicating the current status of the lead," according to the data dictionary. This column's categories seemed to describe a few different lead characteristics, which were not necessarily related with each other. Some values, such as 'Still Thinking' describe a temporary state of the sale, while others such as 'Already a student' are permanent characteristics of the lead that could affect lead score in different ways. It's conceivable that a

lead could fit into more than one of these categories at the same time, but the field doesn't allow multiple values.

Additionally, some values seem like they would disqualify a lead from converting (eg. 'Lost to EINS' or 'Diploma holder (Not Eligible)'), yet still include converted leads. Notably, the most frequent non-null value is 'Will revert after reading the email,' which seems to be a tag for leads in the late stages of a sale, and would of course be associated with higher converted rate (in other words, the sale was already closed, it's just a matter of time). Although this would surely help our model identify converted leads in this dataset, it wouldn't be a helpful feature for assessing leads earlier along in the process. For all of these reasons, I decided to drop this column from the dataset entirely, as it is not appropriate for inclusion in our model.

The data also contained two pairs of related columns, ‘Asymmetrique Profile’ and ‘Asymmetrique Activity,’ each with a Score (out of 20) and an Index, which was an ordered bin (‘Low’, ‘Medium’, or ‘High’) of the score. For the Profile Index, the bin ranges did not seem appropriate for the data. So, I used the associated Profile Score to recode the ranges. Then, I dropped both the Profile Score and Activity Score columns, as this data was already reflected in the Index columns.

The three remaining numerical columns, ‘visits’, ‘visit_time’, and ‘visit_pages’, each described leads' interactions with the website. Rather than imputing with the median, I decided it was more reasonable to assume null values represent 0 (no value recorded if there were no page visits), which is also the mode of each of these columns.

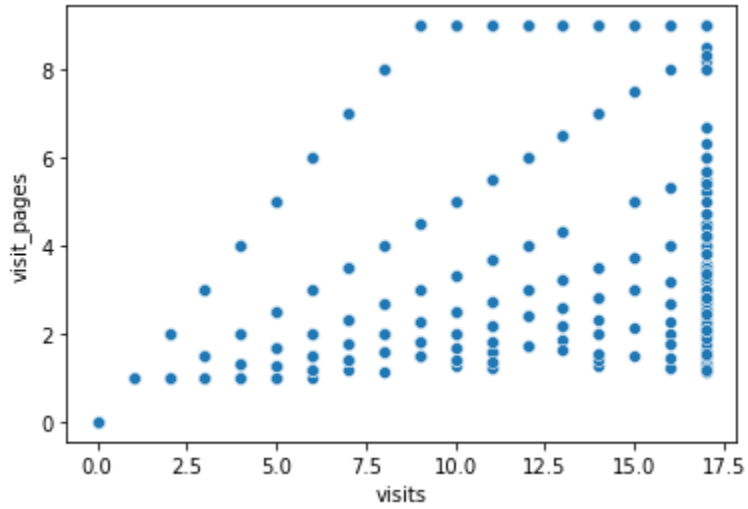
Both ‘visits’ and ‘visit_pages’ had some extreme outliers, though these outliers are most likely legitimate data points. Instead of deleting these rows, I set the value of everything above the 99th percentile to the value of the 99th percentile. These records will still be associated with an abnormally high value, but the absolute scale of the distribution is reduced to a more manageable range.

After these cleaning and feature reduction tasks, the resulting dataset contained all 9,420 original records, with 23 columns (including unique lead number and our target variable, ‘converted’).

Exploratory Analysis

Examination of correlation between columns revealed a few important insights on the data. First, the ‘Unknown’ value of the ‘matters_most’ column was nearly perfectly correlated with having ‘Unknown’ occupation. During the data cleaning phase, I noted that for matters_most, there are only three rows with a value other than "Better Career Prospects" or null. However, I kept the column in case there was any significance to the Unknown (NaN) values. Since the occupation column will also be able to predict any relationship between the Unknowns and our target variable, I decided to drop the matters_most column.

One particularly curious correlation was between visits and visit_pages. According to the data dictionary, visits is "The total number of visits made by the customer on the website," and visit_pages is "Average number of pages on the website viewed during the visits". Yet, as this scatterplot shows, the number of pages seems to be bounded by the number of visits. (When looking at the plot, remember also that the data has also been capped at the 99th percentile)



This relationship seems odd to me -- why should the average number of pages viewed never exceed the number of visits (for example, one visit with 4 pages)? If it were possible, I would seek to clarify what counts as one 'visit', since it looks like each new page visited is being recorded as a distinct visit, when that should not be the case.

Next, I graphed countplots of each categorical column, separating values based on the target variable. By looking at the graphs, some important features were already apparent:

- **Origin**: 'Lead Add Form' has much higher conversion
- **Source**: 'Welingak Website', 'Reference', and 'Unknown' have higher conversion
- **Last Activity**: 'SMS Sent' has much higher conversion
- **Occupation**: 'Working Professional' has higher conversion
- **Quality**: 'Unknown', 'Not Sure', and 'Worst' have low conversion while the other 3 categories have high conversion
- **Profile**: 'Potential Lead' has good conversion
- **Activity and Profile Index**: Higher activity index doesn't seem to result in higher conversion, but high profile index does improve conversion a little
- **Last Notable Action was 'Modified'**: When the last notable action was 'Modified', leads were less likely to convert.

Looking at boxplots of the numerical columns, **Visit Time** tends to be higher for converted leads. I would cautiously say this suggests that if X Education can get potential leads to spend more time on the website per visit, they might be able to increase their conversion rate. There are two important things to remember, though. First, that the directionality of this relationship is

not clear (does spending more time on the website make one more likely to convert, or do leads that are more likely to convert tend to spend more time on the website?). Second, if X Education makes a goal of increasing time spent, they should be careful how they achieve it. Making the website harder to use, for example, could result in more time spent, yet actually have an adverse effect on getting leads to convert.

Modeling

To prepare for modeling, I first used One-Hot Encoding to convert the categorical variables to binary tags for each possible value. As a result, the final X vector for modelling included 139 dimensions. After splitting the data into training and test sets with a comparable percentage of the target group, I fit a standard scaler to the training set in order to standardize the variance of each numerical column. Then, I built and tuned three different models: Logistic Regression, Random Forest, and XGBoost.

For each model I tuned the hyperparameters to optimize the negative log loss. I chose this metric because for the purpose of lead scoring, the predicted probability of conversion is actually more important than accuracy. The purpose of this model is not necessarily to label a lead as 'will convert' or 'will not convert', but instead to give a score out of 10 for how likely they are to convert. The Log Loss function penalizes the model for predicted probabilities that are farther off from the actual value.

After tuning hyperparameters, I compared models based on the area under the ROC curve. The curve is a plot of False Positive Rate vs. True Positive Rate over all possible thresholds of classification. TP Rate is important to our model. We want to make sure that our model finds as many of the converting leads as possible (high recall). We are not as concerned about False Positives, which may be a waste of time for sales staff, but less important than missing possible conversions. But, because we are not locked into one threshold for deciding how to categorize new leads, the ROC curve gives us a better idea of how dependable the lead score will be.

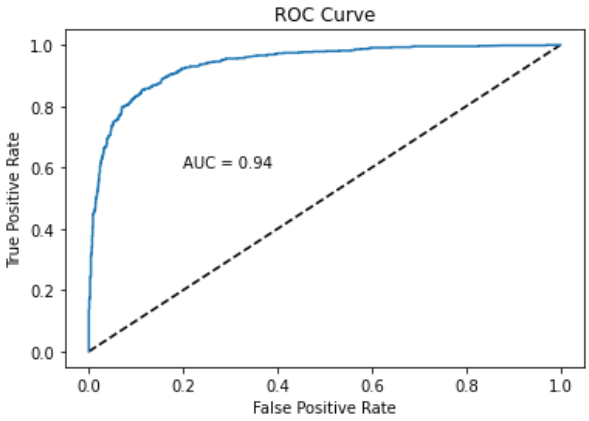
Grid Search Cross-Validation found the best value of C for the LogReg model to be 0.3. The resulting model achieved ROC-AUC of 0.927 when applied to the test set. Recall was 0.78 for positive cases (conversions), and 0.91 for negative cases.

Tuning hyperparameters for the Random Forest model was more difficult, and multiple Randomized Search CV rounds produced varying results. So, I decided to try Bayesian Optimization to identify values that maximized negative log loss. After getting the resulting hyperparameter values from the optimization, I used Randomized CV with smaller ranges around those values, leading to the final parameters: n_estimators=315, max_depth=19, max_features=19.

The Random Forest model slightly outperformed the LogReg model. It had a ROC_AUC of 0.938 on the test set. Recall was 0.79 for conversions, and 0.92 for negative cases.

Finally, I tuned an XGBoost model using Randomized Search CV. The best performance was achieved with 95 estimators, max depth of 4, eta = 0.2, 90% of columns sampled for each tree, and default regularization.

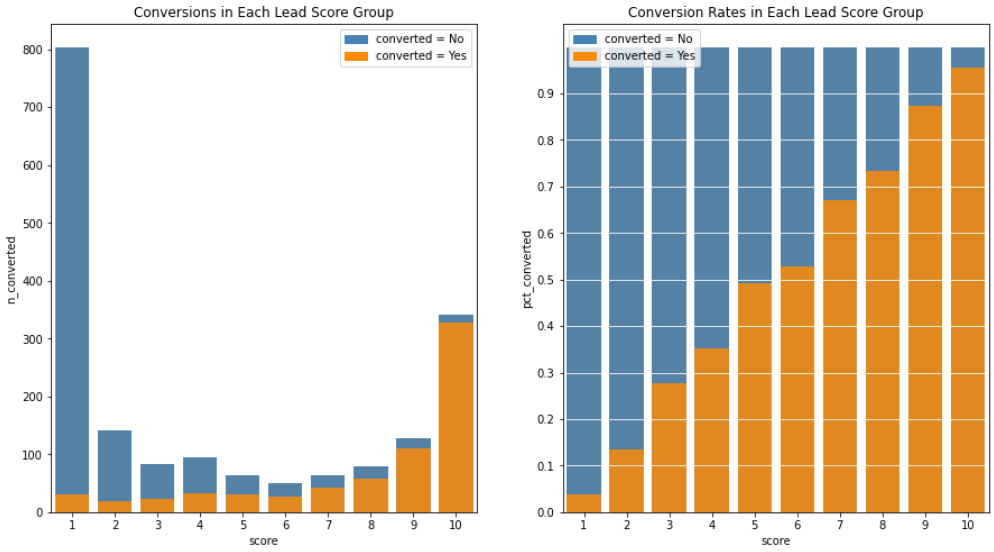
This model produced the best results, with ROC_AUC of 0.94 and recall of 0.81 for conversions.



Interpretation

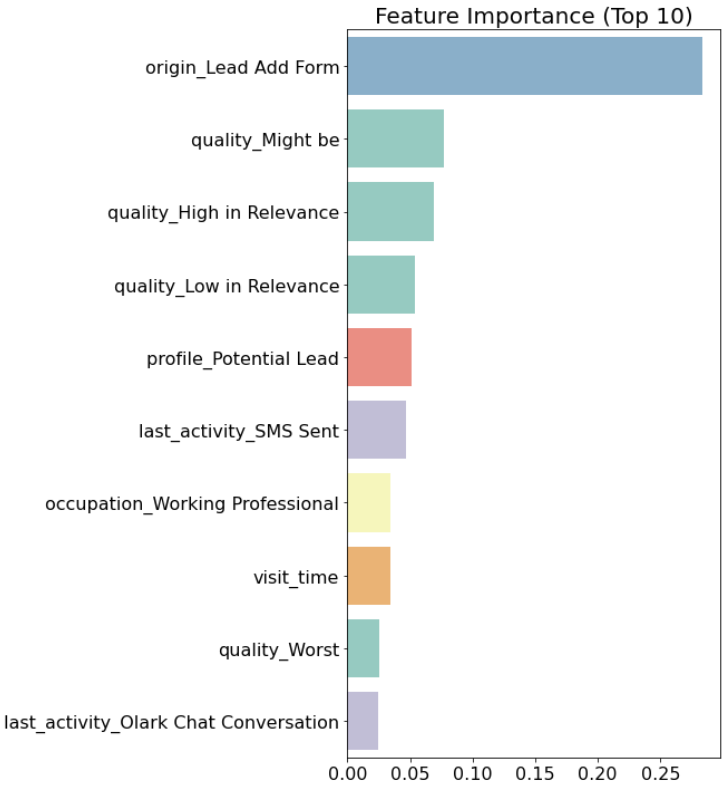
During modeling, I was also careful not to rely too heavily on ROC-AUC and recall alone. The final goal of the model, after all, is not to classify yes/no for conversion, but to assign a probability score. So, another way to evaluate the models is to assign each lead in the test set a score based on the model predicted probabilities, then see what percent of leads with a given score actually converted. Recall that this is why I made sure to optimize log loss, so that the probabilities would be as accurate as possible.

For example, if the model gives 10 different leads a probability of conversion of 80%, we want to see that 8 of those 10 leads actually convert, and 2 do not. The XGBoost model produced the best results in this regard. You can see the predicted vs. actual conversion rates for each lead score group in the test group below.



These graphs make it clear that the model primarily assigns a lead score of 1, meaning it is good at identifying bad leads. Although we might expect the actual conversion rate for these leads to be 10%, the actual conversion rate for this group in the test set was about 3.7%. I argue that this is preferable, based on the expected behavior of our sales staff towards low-scoring leads. In other words, the company is less likely to pursue leads with a score of 1, so the lower the actual conversion in this group, the better.

The model's most important feature was a lead origin of 'Lead Add Form'. Interestingly, this was not the same result as other models, which prioritized Visit Time. The top ten important features in the final model are displayed in this graph, colored by the original column name.



In order to suggest marketing strategies based on these important features, more familiarity with the company procedures and data collection methods is necessary beyond what can be learned from the data dictionary. For instance, to act on the top feature, Lead Add Form, it would be important to know if there is any obvious difference between new leads to fill out the form and leads who do not (origin = API or Landing Page Submission) that the marketing team could pursue.

Another complicating issue is the importance of the 'quality' column, which has four different values included in the top ten most important features. According to the data dictionary, this field "indicates the quality of lead based on the data and intuition of the employee who has been assigned to the lead." So, this is info gathered from sales staff after they've already worked with

the lead. We need to ask at what point in the sales process can we get this data, to make sure that we can score leads before investing too much time.

Working Professionals seem like a great avenue for the marketing team to pursue. During cleaning, I found that Working Professionals only represented about 7.5% of the leads. Since this group tends to have better conversion, it's a market that should be leveraged. As discussed above (see Exploratory Analysis), the Visit Time feature might be useful to improve conversion rate, but the company should be careful about how they attempt to increase average visit time, as some approaches may do more harm than good.

Next Steps

Leadership at X Education should continue to examine the important features identified by the model, and can immediately start adapting their marketing strategy to try to acquire better leads. As mentioned, some industry knowledge will be necessary to determine the exact features that can be leveraged, and in what way they can be pursued.

Meanwhile, the company can apply the scoring model to incoming leads once they have the appropriate data for each lead (including 'quality'). Once they have a score for each group, they should practice different strategies for each group. Again, some industry knowledge as well as trial and error (or carefully designed A/B tests) will help to develop the best strategy for each group. For example, the group of 10s have a very high probability of conversion. Does this mean that they should get special attention? Or, will they tend to convert whether or not they are nurtured by sales staff? Perhaps it is better to pay close attention to leads scoring 5-9, to nudge them towards converting.

Another ongoing strategy to improve outcomes is to expand the available data through further research. I have noted some features that can be dropped or altered to help modeling. Other features that might be helpful include demographic information for the individual leads (age, gender, income, etc.).

Finally, as research, marketing, and sales approaches change, it is important to review the scoring model regularly, because the new strategies may affect the predicted probabilities. Additionally, many other external factors may affect lead conversion rates, such as demand for education, changing costs, changes to competitors, etc. All of these could affect the accuracy of the model. Company leadership would be wise to remember that the model is a guideline to guess at probabilities and guide strategy, not an exact indication of individual lead behavior.