MIT-IDSS 2024

Classification and hypothesis testing

Date – 3/11/2024

Contents / Agenda

- Business Problem Overview and Solution Approach
- Data Overview
- EDA Results - Univariate and Multivariate
- Data Preprocessing
- Model Performance Summary
- Conclusion and Recommendations

Business Problem Overview and Solution Approach

Problem statement :

ExtraaLearn is an initial stage startup that offers programs on cutting-edge technologies to students and professionals to help them upskill/reskill. With a large number of leads being generated on a regular basis, one of the issues faced by ExtraaLearn is to identify which of the leads are more likely to convert so that they can allocate the resources accordingly.

Solution approach/methodology:

As a Data Scientist in ExtraaLearn, My approach will be to build a predictive ML model to predict the probability of the potential lead conversions and allocate better resources to those customers and also loss of opportunity should be minimum.

***Note**: You can use more than one slide if needed.*

Data Overview

- There are total 4612 rows and 15 columns in the data.
- There are no missing values in any column.
- There are no duplicated values in the data as well.
- By observing statistical summary, I found these few key points:

➢ The minimum and maximum age of customers are 18 and 63 respectively and average age of around 46.

➢ On average, customers visited the website 3 times. However, 75 percentile data is 5 while maximum reach up to 30 which suggests that there are outliers in this column.

➢ Average time spent on website is 724 min. And maximum is 2537 min. The median is 376 min while 75%ile is 1336 min which is a significant shift.

➢ In Page views per visit, 75%ile data comes under 3.7 but maximum is 18.4 which indicates there are outliers here.

Data Overview

- Here are some useful observations on categorical columns:

➢ Majority of leads were professional i.e. 56% followed by unemployed.

➢ First interaction via website was more than through app.

➢ People with Low percentage of the profile filled on the website/mobile app were significantly low (107) as compared to high and medium.

➢ Only 233 people have seen ad on newspaper and only 527 on magazine. Also, reference and educational channels were also not so effective to attract the leads. This makes it evident that company is not focusing on better marketing. This shows low outreach of the company on different ad platforms.

EDA Results

- Provide coments on the visualization such as range of attributes, outliers of various attributes.
- Provide comments on the distribution of the variables
- Use appropriate visualizations to identify the patterns and insights
- Key meaningful observations on individual variables and the relationship between variables

❑ **Univariate analysis :**

- The Age distribution looks like bimodal distribution with two peaks. The distribution of people of age 18 to 50 seems uniform around 200.
- The distribution of website visits is right skewed and has outliers on the far side of right whisker.
- Mostly professionals are visiting website more than 10 times as they need more time to make informed decisions or maybe comparing the reviews with others. While, 3 students visited highest number of times(27, 29 and 30 times) who might be highly interested in the services offered. Company should provide additional customization to these high engagement leads.

EDA Results

- Majority of leads spent less than 500 min. As mentioned earlier, there is skewness in the right side of the distribution.

- Mean and median are quite far from each other in time spent on website distribution which makes it clear as some leads are spending more than 1300 min on the website. These may be frequent visitors and genuinely potential leads which has high likelihood of converting. They should be given more focus.

- As mentioned in overview, there are outliers in the page views per visits. The average number of pages visited were 3. Some leads viewed more than 12 pages which maybe the same people who spent highest time on the website.

- Majority of leads are professional i.e. 56.7%.

- First interaction of leads is higher on website than app which is obvious as website is more accessible than app.

EDA Results

- First interaction of leads is higher on website than app which is obvious as website is more accessible than app.

- Almost 98% leads had their profile high or medium completed. This can be a positive sign that people are genuinely interested. We need to evaluate further if these are the leads who visit most often on the website.

- Email activity was highest than phone or website live chat which can be because most of people are professionals and unemployed and they use email more often.

- Only 10% leads had seen the ad on newspaper and 5% on magazine which maybe due to the digitalization of traditional media.

- Around 88% lead had not seen ad on digital media. We should focus on allocating resources for marketing to get broader reach to get higher conversions.

EDA Results

- Only 2% of the leads had heard about ExtraaLearn through reference. This is due to the low marketing and low sense of credibility.

- 30% leads are converted which makes sense as there is low outreach and ineffective counseling of targeted leads who have high chance of conversion.

❑ **Bivariate analysis :**

- Conversion rate on website is positively correlated to time spent on website and pages views per visit have no correlation with status.

- Website visits has slight positive correlation with pages views per visit and time spent on website.

- Age is not correlated much to any of the factors but has slight correlation with conversion.

EDA Results

- Professionals and unemployed had more conversion rate than students. It maybe because professionals and unemployed seek to gain skills to advance their career or get a job. Students, on the other side, don't feel much urgency to learn advanced skills as they are not urgently looking for a job. Or they may have different expectations.

- Students are below 30 years old which is obvious and professional & unemployed range between 25 to 60. As per distribution, students are less in number. They should be given more focus.

- As per observation, professionals and unemployed are more experienced and engaging while students are less interested in the offerings of the company. Emphasis should be also given to providing courses compatible to students so that they can manage the academics with courses. Or partnerships with academic institutions.

EDA Results

- People who first interacted on website had more conversion than on app.
- Distribution plot of time spent on website with status suggest that leads who spent more time on website are more likely to end up buying the courses which I mentioned earlier. Focus should be given to make website more engaging.
- Website visits have no correlation with conversion. Same with the pages views per visit.
- Leads who completed their profile are more interested in buying the courses. On the other hand, leads who left their profile incomplete were not genuinely find motivation to buy.
- Leads who last interacted on website had more conversions than email and phone. It can be because focus is not properly given on the people who left their profiles incomplete.

EDA Results

- Newspaper ad and any media platform has not been effective for improving conversion rates.
- Referral has significantly improved conversion of leads. Some referral points or awards should be given to promote more referrals as people trust referred by friends or relatives.

Model Building

- Provide insights on the performance of different models.
- Provide comments about model performance after tuning the hyperparameter using GridSearchCV.
- Choose the model performance metric and provide reasoning for the same.

❑ **Decision Tree :**

➢ On training data :

- There are 0 errors on the training set. Model performance is excellent.
- Classification of samples is exceptionally perfect which indicate overfitting on the training set.
- Hence lets check the performance of decision tree classifier on the unseen test data.

Model Building

➢ **On test data :**

- On the test data, model's performance is much better for class 0 which has more samples i.e. 962 while it is weaker for class 1(422).
- The precision and recall for class 0 are 87% and 86%. F1 score is 86% which is good performance. For class 1, there is a noticeable drop in the above metrics (precision-69%, recall-70%, F1 score-70%).
- The drop in precision and recall in both classes of target variable is due to the class imbalance. So we will further perform hyperparameter tuning by GridSearchCV to improve the results.

❖ **After Hyperparameter Tuning :**

➢ **On Training data :**

- Low recall means loss of potential customers so we would want recall for class 1 to be maximized. So we performed hyperparameter tuning to minimize false negatives or to maximize recall.
- As we can observe, after tuning model shows higher recall for class 1(88%) and slightly lower recall for class 0(77%).

Model Building

- Precision(94%) for class 0 and 62% for class 1 indicates that there is some imbalance in both the classes. But it performed reasonably good, with f1 score of 73% for class 1 and 85% for class 0 .
- The model after tuning performed better by improving recall, although at the cost of precision.

➢ **On test data :**

- The recall for class 1 has improved for class 1(86%) from before tuning(70%), it means model is achieving a good balance between both the classes. However, there is certain drop in the precision(62%) which was 69% on test data before tuning. So there comes a trade off between precision and recall.
- For class 0, model shows recall of 77% which is notably decreased after tuning, this sacrifice allowed for increasing recall for class 1 which is more important.
- However, there are some differences between precision of class 0 (93%) and class 1 (62%) which could be further improved by ensemble techniques.

Model Building

❑ Random Forest :

➢ On training data :

- As per metrics table, we can see that precision, recall and f1 score are all equal to 1.0 which a perfect score which means it is overfitted to the training set.
- Random forest consist of multiple decision trees and higher selection of features usually overfits on the training dataset, which leads to poor generalization on unseen data.
- Hyperparameters tuning can possibly overcome this overfitting model and improve generalization on validation set.
- Lets check the performance metrics on the test data set.

Model Building

➢ **On the test data set :**

- As we can observe from the metrics table, precision and recall of the random forest model on test set are 87% and 91% respectively for class 0(i.e. lead will not buy). High precision and most importantly very good recall showing better effectiveness of the model in predicting the outcome.

- F1 score of 89% is also showing overall reliable performance for class 0.

- For class 1(i.e. lead will buy), 78% precision shows that there are some false positives in the predictions(82). Recall of 0.68 for class 1 indicates there are many instances of false negatives which we are aiming to minimize in order to avoid loss of opportunity.

- 73% F1 score is significantly lower than other models for class 1 which is mostly affected by class imbalance. Lower f1 score maybe due to major trade off between precision and recall

Model Building

- 73% F1 score is significantly lower than other models for class 1 which is mostly affected by class imbalance. Lower f1 score maybe due to major trade off between precision and recall for class 1 samples.

- Due to imbalance in samples in the classes, model showed biasness in performance for class 0 than class 1. Like decision tree, hyperparameters tuning can further enhance the overall metrics scores for class 1, especially recall.

❖ **After Hyperparameter tuning :**

➢ **On training data :**

- For class 0, we can observe that precision increased with 94% while recall decreased to 83% as before tuning, this maybe due to trade off to improve class 1 recall.

Model Building

- F1 score for class 0 have not improved much but 88% is overall a good performance for class 1 instances predictions.

- For class 1, recall has quite significantly raised to 87% at the expense of precision which indicates tuning model has proven to be effective in minimizing false negative cases in the predictions.

- F1 score of class 1 has also increased slightly which is better score for evaluating the predictive power of the model after tuning.

➢ **<u>On test data :</u>**

- The metrics score for both class 0 and class 1 are almost similar to training set. This has made clear that model generalizes better on unseen data, which is best model to predict the further possibilities. This seems the best model as compared to other models as it shows an excellent generalization on unseen data.

Feature Importance

- **Time_spent_on_website**(0.34814) and **first_interaction_website**(0.32718) are the most important features. This implies that online engagement is important factor to determine the conversion rates of potential leads.
- Another important feature is **profile_completed_medium** with importance level 0.23927, indicates that leads who have completed their profiles with medium level are more interested and likely to convert more.
- **Age** with 0.06389 is notably less predictive than above top features.

Feature Importance

❑ Decision Tree :

- Other features like **website_visits, page_views_per-visit, last_activity_Phone** have zero importance, which we have observed in bivariate analysis of these features with status, we found these features don't contribute in conversion of customers.

- Media platforms like **print_media_type1_yes, print_media_type2_yes, digital_media_yes** and **educational_channels_yes** also have importance score of zero. It makes it clear that marketing has not been effective in converting the leads. It needs better guidance and resources to improve the marketing strategies.

- **Current_occupation_student** and **current_occupation_unemployed** are not helping much for converting which we have already seen during bivariate analysis. These features are considered as weak predictors by the model as these features might not be effective on conversion.

Feature Importance

- Overall, model suggesting online engagement metrics to be the key elements in predicting whether the lead will convert or not. These features are influencing the behavior pattern of the leads. Other factors like newspapers, magazine and educational channels have negligible effect on lead's possibility of expected outcome. This is indicating that better optimization of digital platforms could increase the likelihood of more conversions in the future.

Feature Importance

❑ **Random Forest classifier :**

- Similar to the decision tree model, **time spent on website, first_interaction_website, profile_completed, and age are the top four features** that help distinguish between not converted and converted leads.

- Unlike the decision tree, **the random forest gives some importance to other variables like occupation, page_views_per_visit, as well.** This implies that the random forest is giving importance to more factors in comparison to the decision tree.

Model Performance Summary

- Here is the Model performance summary of all the models :

	Class 0 precision	Class 0 recall	Class 0 F1 score	Class 1 precision	Class 1 recall	Class 1 F1 score
Decision Tree(untuned)	0.87	0.86	0.86	0.69	0.70	0.70
Decision Tree(Tuned)	0.93	0.77	0.84	0.62	0.86	0.72
Random forest(untuned)	0.87	0.91	0.89	0.78	0.68	0.73
Random forest(Tuned)	0.93	0.83	0.87	0.68	0.85	0.76

Model Performance Summary

- After observing the table, we can conclude that random forest model(tuned) performed the best among all other models. The model achieved a good balance between precision and recall for both classes.

- Recall for class 1 in this model is highest which suggests that it predicted less false negatives.

- As it performs the best, it can be the most reliable model among other models.

APPENDIX

Slide Header

- Please add any other pointers or screenshots (if needed)

Great Learning

Happy Learning !