

LEAD SCORING

By Karlotta Kürzel





PROBLEM STATEMENT

- X Education, an online course provider, faces a challenge of low lead conversion rates despite a substantial influx of leads from various sources, including website visitors, forms, and referrals. While the company actively engages potential customers through sales efforts, the current lead conversion rate of approximately 30% is below the desired level. The company aims to improve lead conversion efficiency by identifying and prioritizing the most promising leads (so called hot leads), thus increasing the conversion rate to meet the target of 80%.

BUSINESS GOAL

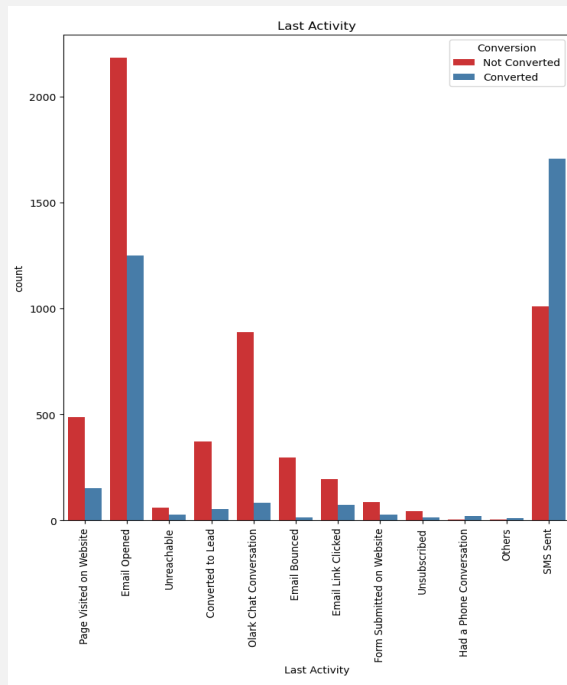
- The primary objective of X Education is to enhance its lead conversion process and achieve a higher conversion rate of 80%.
- Additionally, a lead score between 0-100 should be assigned to each lead in order to identify hot leads
- The company seeks to optimize its resources by focusing on leads with the highest potential for conversion, enabling the sales team to engage in more effective communication strategies.
- By implementing a logistic regression model, X Education aims to assign lead scores that reflect the likelihood of conversion, allowing the sales team to prioritize leads and allocate resources efficiently.

ANALYSIS APPROACH

- Import and explore data
- Clean data (e.g., null value treatment, dropping of unnecessary columns, outlier treatment)
- EDA
- Prepare data (e.g., dummy variable creation, train/test split, scaling)
- Build logistic regression model (RFE and VIF)
- Model evaluation
- Test model on test data and evaluate
- Assigning a Score between 0-100 to each lead

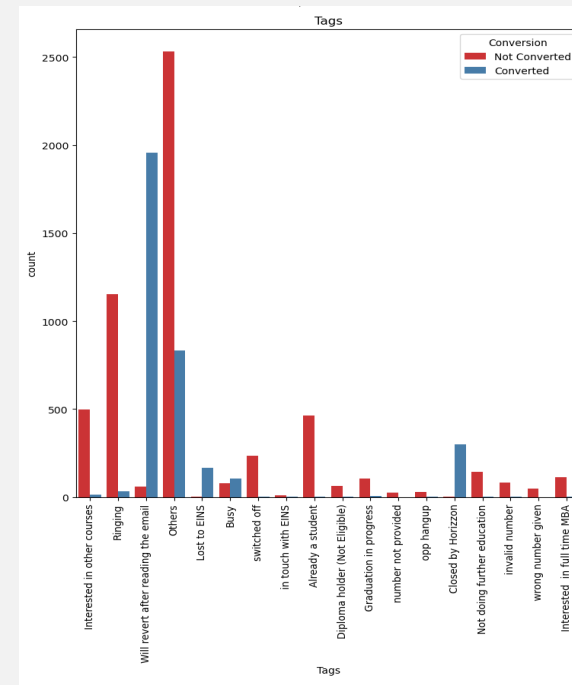
EDA

In the following, uninformative/ highly skewed columns were removed prior and low frequency values were summarized under a category “Others” if applicable



Last Activity

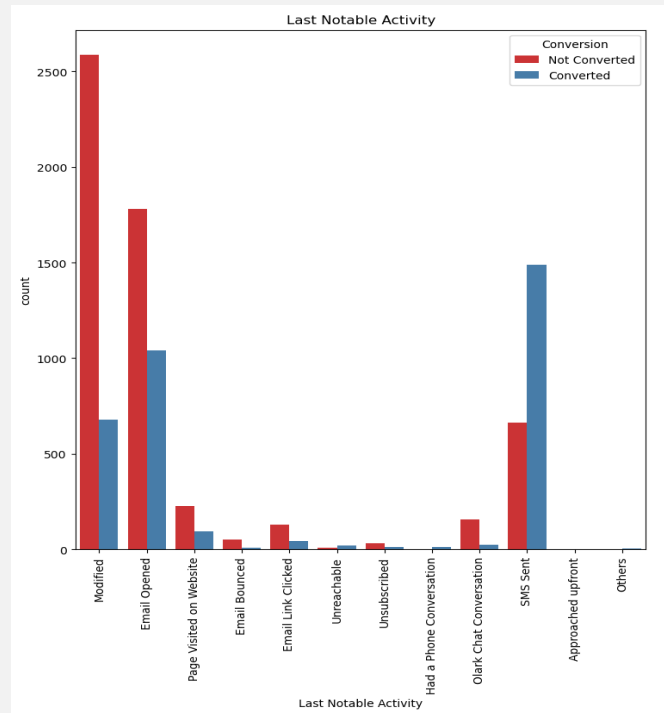
- Most people fall into the category Email opened
- SMS Sent and had a Phone Conversation have high Lead conversion



Tags

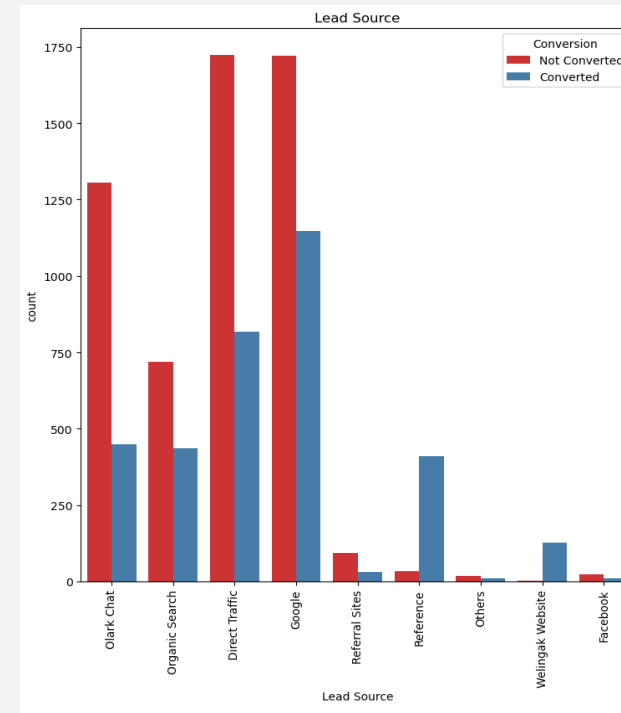
- Will revert after reding the email is the most frequent category
- Will revert after reding the email, Lost to EINS, Busy and Closed by Horizon have high Lead conversion

EDA



Last Notable Activity

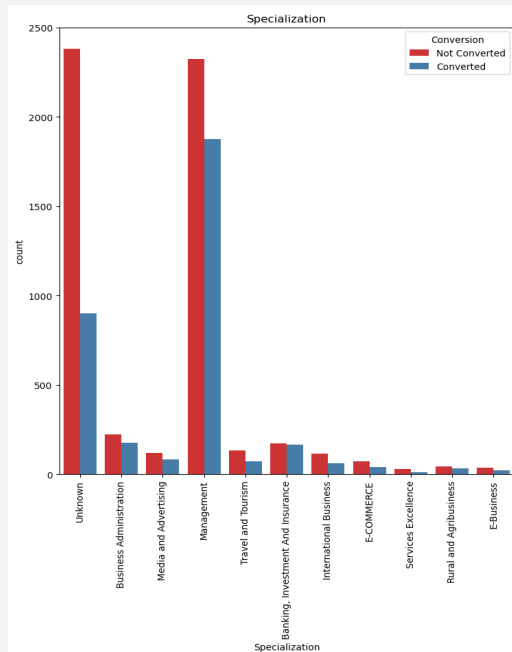
- Most leads fall into the category modified or email opened
- SMS Sent has a high Lead conversion



Lead Source

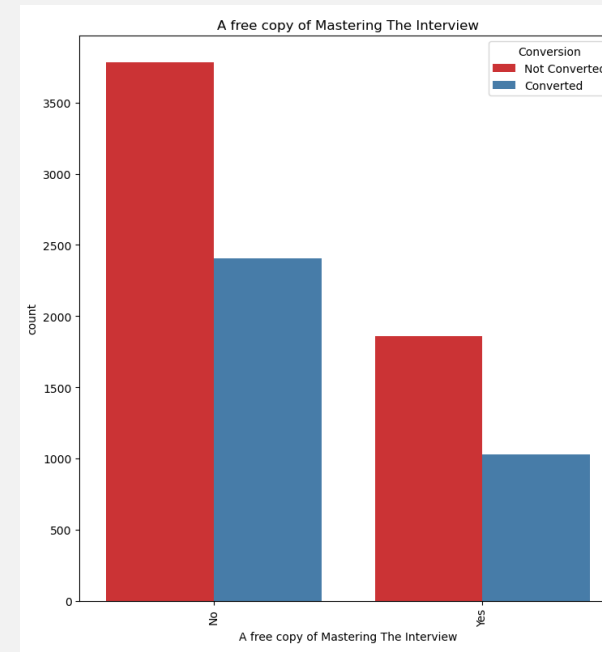
- Most leads come from direct traffic or google
- Reference and Welingak Website have high Lead conversion

EDA



Specialization

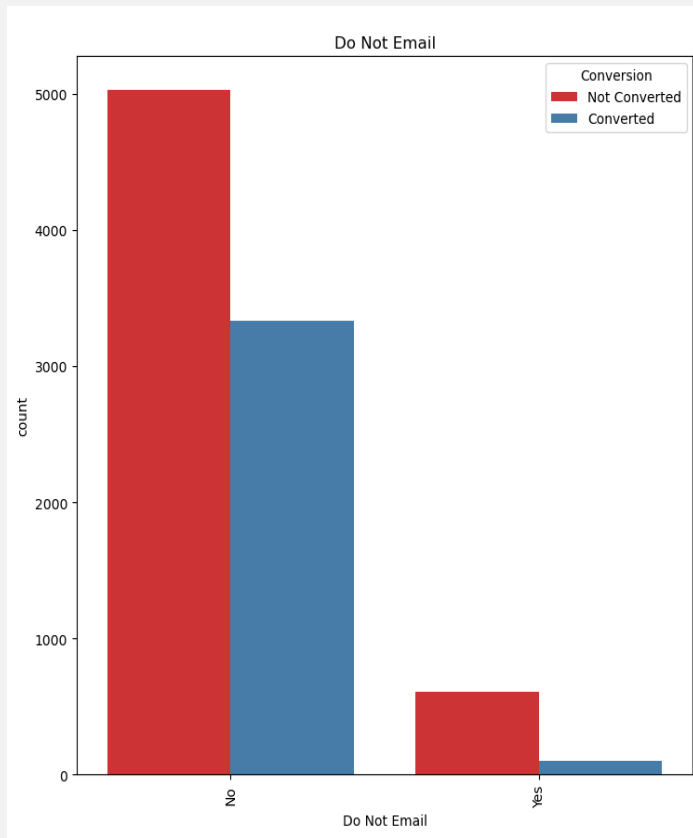
- Most leads either have a management or an unknown specialization
- Specialization does not appear to contribute to high Lead conversation



A free copy of Mastering the Interview

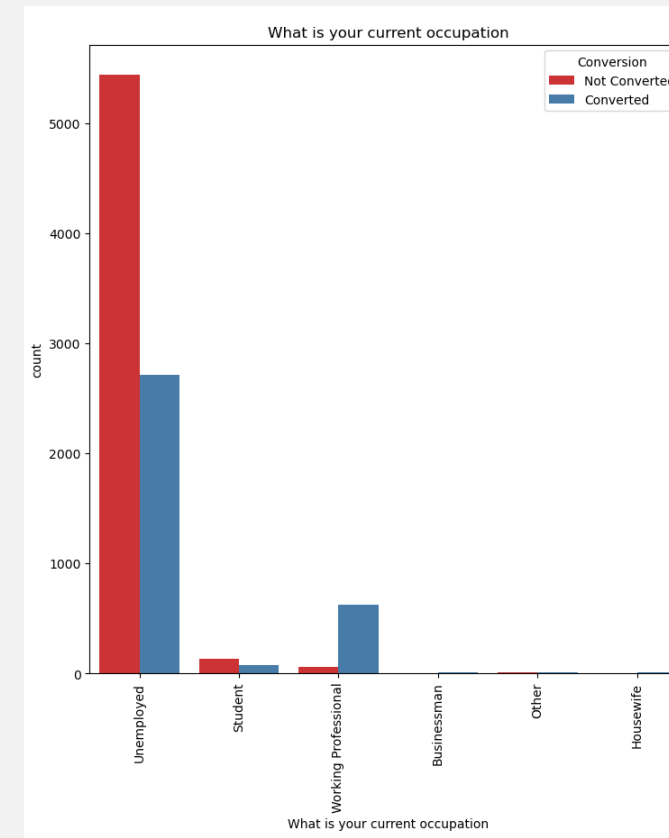
- Most leads do not take a free copy of Mastering the interview
- The category does not appear to contribute to high Lead conversation

EDA



Do not email

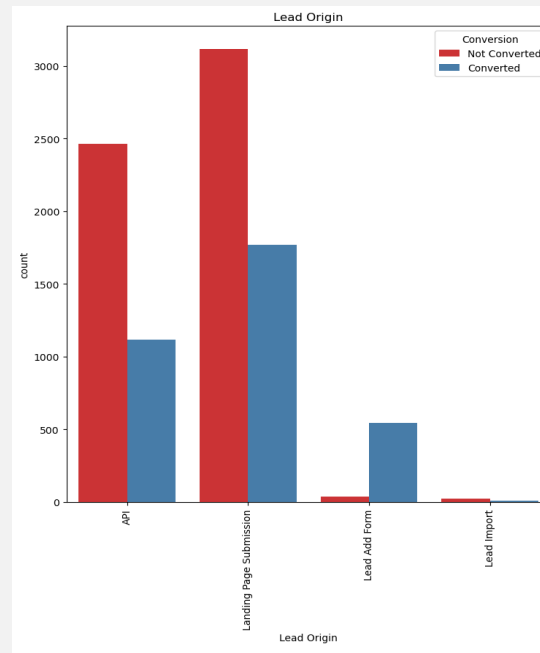
- Most people prefer to be contacted via email
- does not appear to contribute to high Lead conversation



What is your current occupation

- Most leads are unemployed
- Working Professional has a high Lead conversation

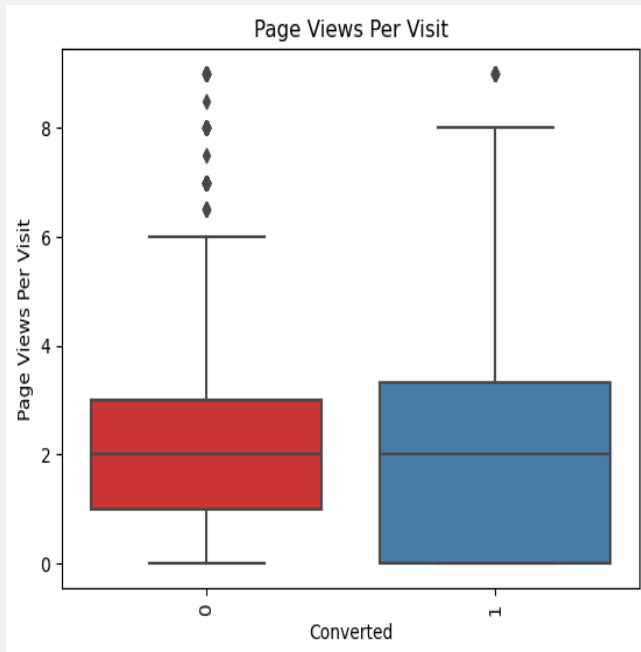
EDA



Lead Origin

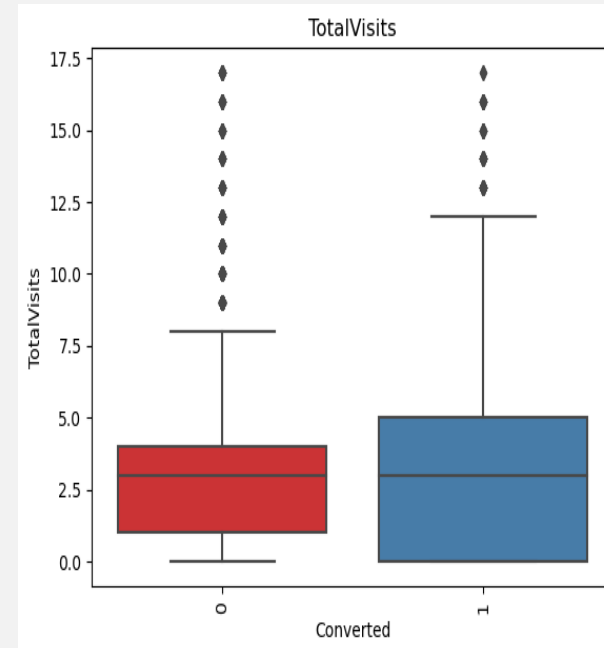
- Landing Page Submission is the most frequent category
- Lead Add Form has a high Lead conversion rate

EDA



Page Views Per Visit

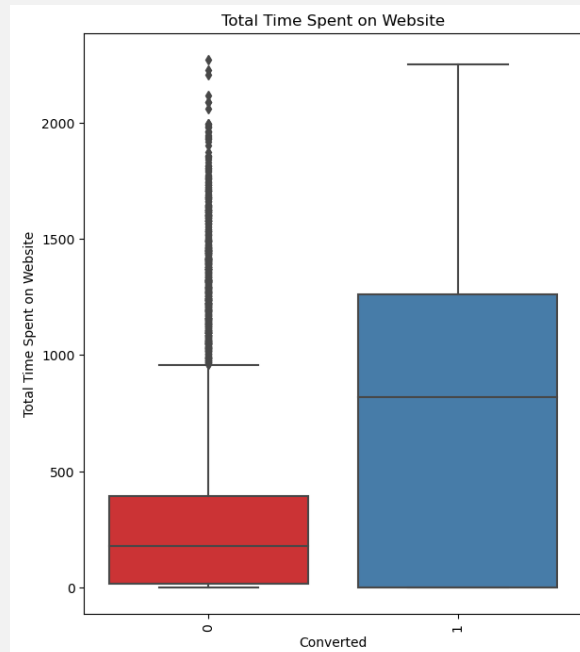
- The median between converted and non converted is very similar



Total Visits

- The median between converted and non converted is very similar

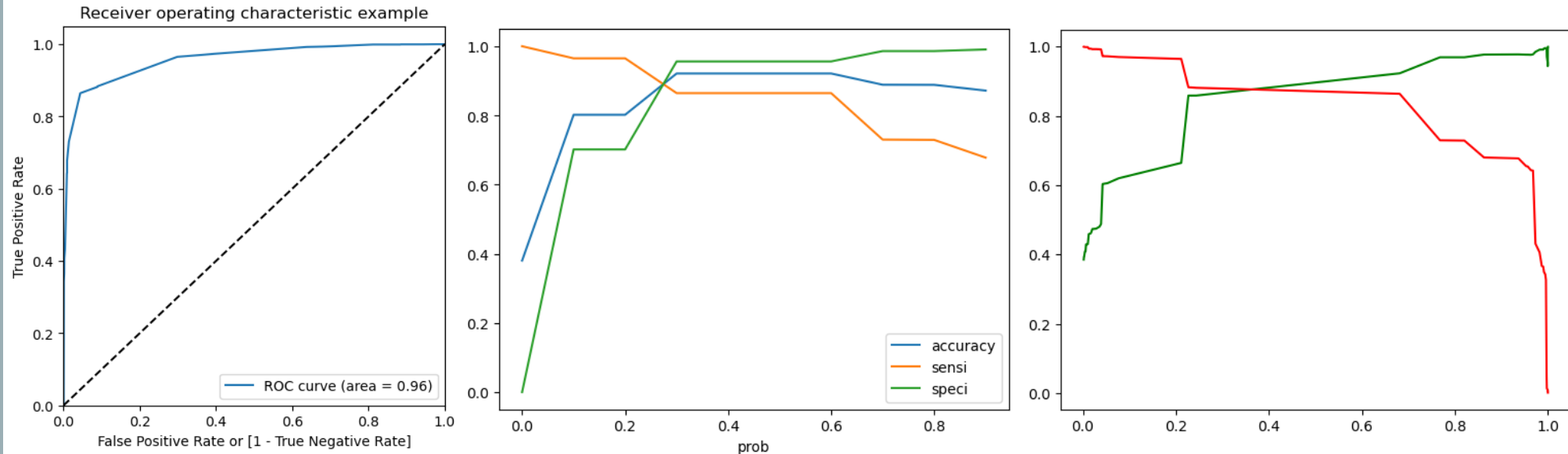
EDA



Total Time Spent on Website

- Converters spent more time on the website than non-converters

TRAINING MODEL EVALUATION



- The ROC curve shows a favourable graph (based on the training set).
- The optimal cutoff is at around 0.28 based on accuracy, sensitivity and specificity (based on the training set).
- 0.38 is the optimal cutoff based on precision and recall (based on the training set).

FINAL MODEL

TRAIN

ACCURACY 91.9%
SENSITIVITY 86.4%
SPECIFICITY 95.2%

TEST

ACCURACY 92.8%
SENSITIVITY 87.1%
SPECIFICITY 96.1%

- The final model encompassed 12 features alongside a constant
- $\text{Logit}(P) = -1.3352 + 7.2024 * \text{Tags_Closed by Horizzon} + 5.8929 * \text{Tags_Lost to EINS} + 4.7345 * \text{Tags_Will revert after reading the email} + 4.4264 * \text{Lead Source_Welingak Website} + 2.0169 * \text{Last Activity_SMS Sent} - 1.5555 * \text{Last Notable Activity_Olark Chat Conversation} - 1.8582 * \text{Last Notable Activity_Modified} - 1.9641 * \text{Tags_Interested in other courses} - 2.7328 * \text{Tags_Already a student} - 3.1794 * \text{Tags_Ringing} - 3.8167 * \text{Tags_invalid number} - 4.2308 * \text{Tags_switched off}$
- A lead score between 1-100 was assigned to each lead during the process

CONCLUSION AND RECOMMENDATION

- Based on the evaluation metrics, the model performance is very good
- Categories such as Tags_Closed by Horizzon, Tags_Lost to EINS, and Tags_Will revert after reading the email positively influence conversion rates.
- Negative effects on conversion rates arise especially from Tags_switched off, Tags_invalid number, and Tags_Ringing.
- Leads interested in other courses or already students should be avoided in the calling process but working professionals or such with Lead Source Reference are promising.
- Considering that converters have a higher total time spent on the website, it is advisable to focus on creating an engaging and captivating website
- Implementing automated email and SMS systems can enhance conversion rates by conserving manpower.
- Consider reallocating resources from Olark Chat Conversation, which negatively affects conversion.