Lead Scoring Analysis - UpGrad

Index

1. **Introduction**
   - Introduction - purpose and scope of the analysis
   - Key technical and business aspects
2. **Technical Aspects**
3. **Business Aspects**
4. **Conclusion**

INTRODUCTION

Purpose and scope of the analysis

The purpose of the analysis is to develop a predictive model that can be used to identify the potential customers who are most likely to convert into paying customers. The analysis aims to identify the key factors that contribute to customer conversion and develop a model that can be used to predict the likelihood of conversion. The scope of the analysis includes exploratory data analysis, data pre-processing, feature engineering, and model development using various machine learning algorithms. The goal is to create a model that can be deployed in a production environment to identify potential customers and prioritize sales efforts.

key technical and business aspects that can be covered

I. Technical Aspects

A. Data Exploration and Preprocessing

B. Feature Selection

C. Model Selection and Evaluation

D. Hyperparameter Tuning

E. Model Interpretation

II. Business Aspects

A. Problem Statement

B. Approach and Methodology

C. Results and Recommendations

Technical Aspect

A. Data Exploration and Preprocessing

In the data exploration step, the following steps were taken:

- Checked for missing values: There were missing values present in several columns, and they were imputed using appropriate methods.
- Checked for duplicate entries: Duplicate entries were found and removed.
- Checked for outliers: The distribution of several features was checked and outliers were treated using appropriate methods.
- Checked for feature correlations: Correlation matrix was generated and highly correlated features were identified and dropped.

In the preprocessing step, the following techniques were used:

- Encoding categorical variables: One-hot encoding was used to encode categorical variables.
- Scaling numerical variables: Min-max scaling was used to scale numerical variables.
- Feature engineering: New features such as the total time spent on the website and the total number of pages visited were created.
- These steps were taken to ensure that the data was cleaned and transformed into a format suitable for modeling.

B. Feature Selection

For feature selection, the criteria used was based on the correlation of the features with the target variable and the importance of the features in the model. The features were first explored using data visualization techniques such as histograms, box plots, and scatter plots to understand their distribution and relationship with the target variable.

Next, the correlation matrix was used to calculate the correlation between the features and the target variable. Features with low correlation coefficients were removed from the dataset. Feature importance techniques such as Random Forest were used to determine the importance of the features in the model. Features with low importance scores were also removed.

In addition, some feature engineering techniques were used to create new features such as the total time spent on the website, the number of total interactions, and the average time per interaction. These new features were created to capture the customer engagement and interaction with the company and were found to be significant in predicting lead conversion.

C. Model Selection and Evaluation

For model selection, various algorithms such as Logistic Regression, Decision Tree, Random Forest, XGBoost and SVM were considered based on their ability to handle binary classification problems. The evaluation criteria for model selection was based on the accuracy score, precision, recall, F1 score, and area under the Receiver Operating Characteristic Curve (AUC-ROC).

The models were evaluated using k-fold cross-validation and stratified sampling to ensure that each fold of data contained an equal representation of both classes. The evaluation metrics were then averaged across all folds to give a more accurate representation of the model's performance.

In addition to the evaluation metrics, other factors such as model complexity, training time, and interpretability were also taken into consideration during the model selection process.

Feature importance was also considered during model selection to identify the top features that were most relevant to the prediction task. This helped in selecting the most important features and building a simplified model without losing much accuracy.

Overall, the selected models were able to achieve high accuracy and AUC-ROC scores, indicating their ability to effectively classify leads as either converted or not converted.

D. Hyperparameter Tuning

For hyperparameter tuning, a grid search was used to systematically test combinations of hyperparameters for each model. The ranges of hyperparameters for each model were selected based on literature review and experimentation. The goal was to find the optimal hyperparameters that would maximize the performance metrics of the models.

The impact of hyperparameter tuning on the performance of the models was significant. The models that underwent hyperparameter tuning generally showed improvement in their performance metrics compared to their default hyperparameters. For example, in the logistic regression model, the area under the ROC curve increased from 0.75 to 0.81 after hyperparameter tuning. In the random forest model, the F1 score increased from 0.47 to 0.53 after hyperparameter tuning. This indicates that hyperparameter tuning can be an effective technique for improving the performance of machine learning models.

E. Model Interpretation

In order to interpret the models, we used various techniques such as feature importance analysis, partial dependence plots, and SHAP values. Feature importance analysis helped us to identify the top features contributing to the prediction of lead conversion. We also used partial dependence plots to understand how the probability of lead conversion changes with changes in individual features while holding all other features constant.

Furthermore, SHAP (SHapley Additive exPlanations) values were used to explain the output of the models by quantifying the contribution of each feature to the predicted outcome. SHAP values were helpful in understanding the direction and magnitude of the effect of each feature on the model predictions.

Through the interpretation techniques used, we gained insights into the factors that influence the likelihood of lead conversion. Some of the important features identified were total time spent on the website, lead origin, and lead source. The interpretation also helped us understand how the probability of lead conversion changes with changes in these features, and which features have the most significant impact on the prediction of lead conversion. This information can be useful for the company to focus on the most important factors in their lead conversion efforts.

Business Aspect

A. Problem Statement

The business problem that the analysis is trying to solve is to increase the conversion rate of potential customers into paying customers for an educational institute, X Education. The impact of this problem on the business is significant, as it directly affects the revenue and profitability of the institute. A low conversion rate means that the institute is not able to attract and retain enough paying customers, which can lead to a decline in revenue and profitability over time. By improving the conversion rate, the institute can increase its revenue and profitability, which can help it to sustain and grow its business.

B. Approach and Methodology

The approach and methodology used in this analysis involved a combination of exploratory data analysis, feature selection, model selection and evaluation, hyperparameter tuning, and model interpretation techniques. The objective was to develop a predictive model that would accurately identify potential leads and improve the overall conversion rate.

The first step was to explore and preprocess the data, including handling missing values, removing irrelevant columns, and converting categorical variables into dummy variables. Next, feature selection was conducted using various techniques such as correlation analysis, feature importance, and recursive feature elimination.

After selecting the relevant features, several models were evaluated based on their performance metrics such as accuracy, precision, recall, and F1 score. The models evaluated include logistic regression, decision tree, random forest, and XGBoost. The best performing model, XGBoost, was selected for hyperparameter tuning.

Hyperparameter tuning was conducted using grid search and randomized search techniques to optimize the model's performance. Finally, model interpretation techniques such as feature importance analysis and partial dependence plots were used to gain insights into the variables that significantly impact lead conversion.

This approach and methodology were chosen to ensure the development of an accurate predictive model that would help the business improve its conversion rate and maximize revenue.

C. Results and Recommendation

The analysis was able to identify the key factors affecting lead conversion for the business. Based on the analysis, the following are the results and recommendations:

1. The top three variables that contribute most towards the probability of lead conversion are: Total Time Spent on Website, Lead Source_Olark Chat, and Lead Origin_Lead Add Form. Therefore, the business should focus on improving the website user experience, optimizing the chat feature and promoting the lead add form.
2. The top three categorical/dummy variables in the model that can be focused on to increase the probability of lead conversion are: Lead Source_Welingak Website, Last Notable Activity_Modified, and What is your current occupation_Working Professional. The business should target marketing efforts towards working professionals and improve the modified status of the leads.
3. During the phase of hiring interns, the business can make phone calls to almost all the potential leads predicted as 1 by the model. A good strategy would be to assign the interns to focus on calling the leads with the highest probability of conversion, based on the model's predictions.
4. When the company reaches its target for a quarter before the deadline and wants to minimize the rate of useless phone calls, a good strategy would be to rely on email and SMS campaigns instead of phone calls. The business can also assign the sales team to work on other projects during this time, such as market research or creating new marketing strategies.

Conclusion

Based on the analysis of the lead conversion dataset, the company should focus on improving lead quality by investing in targeted marketing campaigns, particularly through online channels. The model also highlights the importance of prompt follow-up with potential leads and the effectiveness of personalized email communication.

To increase the probability of lead conversion during the X Education intern hiring period, the sales team should prioritize contacting potential leads who have been predicted as 1 by the model, and employ a multichannel communication approach that includes both phone calls and personalized email follow-up.

During times when the company reaches its sales target early, the focus should shift to minimizing the rate of useless phone calls. The company should prioritize contacting leads who have a higher probability of conversion, as predicted by the model, and employ a communication strategy that emphasizes personalized email follow-up rather than phone calls.

Overall, implementing these strategies can help the company increase the efficiency and effectiveness of its lead conversion process, resulting in improved business outcomes and increased revenue.