Retail Customer Classification And Email Marketing Optimization

By Vijay Sithambaram

Introduction

Introduction:

- This project is centered around the creation of a sophisticated customer classification model tailored to predict various outcomes, with a specific focus on unraveling intricate patterns associated with email marketing opt-in behavior.
- The accurate anticipation of these outcomes is pivotal for orchestrating well-informed marketing strategies, ultimately enhancing customer engagement and retention rates.

Data Overview

Dataset Details:

- Our dataset comprises a comprehensive collection of over 52,000 sales transactions, spanning across a diverse spectrum of more than 12 distinct departments.
- Imprinted within this expansive dataset are a plethora of unique customer attributes, ranging from demographic information such as age, post code, and gender to multifaceted purchasing behaviors.
- Temporally, the dataset encapsulates a substantial period, extending from the commencement of 2019 to the early months of 2021, thereby encapsulating approximately two years' worth of intricate transactional chronicles.

Feature Engineering

Key Features for Classification:

- Customer Duration:
  - Imposingly quantified as the temporal interval between a customer's inaugural and concluding transactions, this pivotal feature has been meticulously computed.
  - With a mean duration of 240 days and a median duration of 213 days, this metric offers profound insights into individual customer lifecycles.
- One-Hot Encoding:
  - To encapsulate the nuances of categorical variables, a rigorous one-hot encoding regimen was judiciously applied to attributes such as post code and the frequencies of transactions within distinct departments.
  - This process has led to the generation of binary features, which in turn significantly enriches our model's interpretability.
- Standardization without Mean:
  - Prudent standardization of the dataset was judiciously executed, sans the subtraction of the mean. The rationale underpinning this approach is to retain the innate distributional characteristics of our data.
  - Leveraging this method, we safeguard against undue information loss while expediting model convergence during the rigorous training phase.

Addressing Class Imbalance

- Class Imbalance:
  - The prevailing class distribution exhibits a prominent skew towards the opt-in category, which constitutes approximately 85% of instances, while the opt-out category constitutes the minority, accounting for the remaining 15%.
- Approach to Address Imbalance:
  - Upsampling Minority Class:
    - Mitigating the deleterious impact of class imbalance necessitated the judicious upsample of instances belonging to the opt-out class, thereby rectifying the class distributional incongruity.
    - By meticulously amplifying the opt-out class instances through replacement, we have attained a harmonious equilibrium wherein the opt-in and opt-out categories are evenly represented at approximately 50% apiece.

Model Selection And Comparison

- Models Explored:
  - Naive Bayes (Baseline):
    - As an initial point of reference, the Naive Bayes algorithm was employed, serving as a foundational probabilistic classifier for benchmarking purposes.
  - Logistic Regression:
    - The Logistic Regression model, characterized by its linear nature, was enlisted to capture discernible linear associations amongst the diverse array of features at play.
  - Decision Tree & Random Forest:
    - Introducing complexity, the Decision Tree and Random Forest models operate adeptly in unearthing intricate non-linear patterns etched within the data.
  - K-Nearest Neighbors (KNN):
    - Functioning on a proximity-based principle, KNN is an instance-based algorithm adept at identifying localized patterns.
  - Support Vector Classifier (SVC):
    - The hyperplane-conforming Support Vector Classifier excels at discerning optimal class boundaries, thereby manifesting pronounced class separation.
  - XGBoost:
    - An ensemble technique, the XGBoost algorithm harnesses the power of gradient boosting to holistically capture complex and interrelated relationships.
- Methodology for Comparison:
  - Evaluation Metric: Accuracy.
  - Cross-Validation:
    - Relying on a robust 5-fold cross-validation regimen, we ensure the equitable evaluation of each model's performance across diverse subsets of the dataset.

Model Performance And Conclusion

- Final Model Selection: Tuned Support Vector Classifier (SVC):
  - Elevating its training accuracy to 90% and test accuracy to approximately 96%, the Tuned Support Vector Classifier (SVC) rightfully claims the mantle of our chosen model.
- Impact:
  - Augmented Predictive Prowess:
    - The SVC emerges as the quintessential champion, evincing a propensity for accurate and nuanced predictions of diverse customer outcomes.
  - Strategic Marketing Informatics:
    - Enriched with actionable insights into email marketing opt-in behavior, our model constitutes a potent tool for the strategic orchestration of targeted marketing initiatives.
- Results and Outcome:
  - Validation of Accuracy:
    - Upon rigorous testing on previously unseen data, the SVC successfully affirms its exceptional generalizability and predictive efficacy.
  - Optimized Marketing Efficacy:
    - Armed with the informative discernments garnered from our model, our marketing strategies stand primed for optimization, inducing a favorable ripple effect across customer engagement paradigms.

Out-Of-Sample Testing

- Testing Tuned Models:
  - Our tuned models undergo rigorous testing on a pristine and uncharted test dataset, thereby mimicking real-world operational dynamics.
- Utilization of ROC-AUC Scores:
  - The venerable ROC-AUC metric is judiciously invoked to furnish us with a holistic depiction of our models' performance across a range of sensitivity and specificity thresholds.
- Test Data Results:
  - Logistic Regression: Racking up a commendable ROC-AUC score of 62.5%.
  - K-Nearest Neighbors (KNN): Scaling upwards with an elevated ROC-AUC score of 75%.
  - Random Forest: Radiating strength, our model proclaims a substantial 95.83% ROC-AUC score.
  - XGBoost: Sustaining its competitive stance, XGBoost showcases a praiseworthy ROC-AUC score of 87.5%.
  - Support Vector Classifier (SVC): Championing excellence, our SVC culminates with an impressive ROC-AUC score of 95.83%.

Future Steps And Conclusion

- Model Performance Recap:
  - SVC and Random Forest: Standouts with the highest ROC-AUC scores.
  - Exhibiting impeccable mettle, these models convincingly herald an era of refined classification paradigms.
- Future Steps:
  - Ensemble Techniques Exploration:
    - Delving further, we intend to navigate the realm of ensemble methods and model stacking, potentially fortifying accuracy to unprecedented echelons.
  - Class Balancing Augmentation:
    - By delving into the expansive pantheon of class balancing techniques, we anticipate an even-handed treatment of classes, nurturing fairness.
  - Feature Engineering Elucidation:
    - Immerse ourselves in the boundless arena of feature engineering, thereby unfurling novel dimensions that bear the potential to augment our model's predictive potency.
- Conclusion:
  - Inception of Profound Insights:
    - With the successful fruition of our customer classification model, we have birthed a portal to the unraveling of customer propensities towards email marketing.
  - Strategic Confluence:
    - The practical ramifications cascade into an era of optimized email marketing campaigns and an astute understanding of customer preferences, culminating in heightened engagement and incisive customer stewardship.