Rodrigo Rodriguez

07/10/2024

Applied Data Science

DSC680 (SUMMER)

**Automating Online Cookies Classification to Mitigate Legal Risk**

**Potential Audience Questions**

1. **What cookies were affected by class imbalances?**
2. Targeting Cookies (12%) and Functional Cookies (8%) were underrepresented in the dataset, while Essential Cookies (60%) and Analytics Cookies (20%) were adequately represented.
3. **Why are certain cookies underrepresented in the dataset?**
4. Targeting and Functional cookies were underrepresented due to the low counts of these cookies currently on our primary domain. Our primary domain was used to build the dataset, so the distribution of cookie categories reflects the type of cookies active on our site.
5. **How was the data split between the training and test sets?**
6. The training and test dataset were split using stratified sampling. This ensures that no single category is absent from the test sets and is adequately represented. This technique works with the data that is available so if a cookie count is low in the overall dataset, it will also be low in the test set.
7. **If all models performed about the same, why is Logistic Regression recommended over the other evaluated models?**
8. Logistic Regression was recommended because it outperformed Random Forest and can be further optimized by tuning hyper parameters, which Naïve Bayes has a very limited number of parameters to tune.
9. **Why were so many variables dropped during data preparation?**
10. Variables that were drop did not add any predictive value. For example, the “Cookie ID” variable is a contextual variable that does not improve the model’s predictive capabilities
11. **Why is F1-Score the primary metric to focus on in the Classification Report?**
12. All metrics in the report are valuable but we focus on F1 because it is a combination of Precision and Recall. F1 can be used as a summary metric without having to go into the details of Precision and Recall.
13. **Why did text need to be processed before modeling?**
14. Processing text data narrows down the characters within the text and retains only the core message that the text conveys. The processed text is then converted to dummy-variables in a process called text-vectorization. The text reduction done during text processing is needed to improve computational load and modeling performance.
15. **Did exploratory data analysis reveal any other insightful information?**
16. Yes, EDA revealed other information like the distribution of text length in the description variable and the distribution of cookies by cookie host. Text distribution is right skewed with most cookies having between 50-250 words. Lastly most cookies are hosted by our web analytics platform and security platforms.
17. **Would this model be required after all cookies are categorized?**
18. Once all cookies are categorized the model would still be an asset that we could leverage to quickly categorize any cookies implemented in the future.
19. **Do we still need a cookie management team after implementing this model?**
20. Yes, the model is an asset that can be used by the cookie management team. However, the team is still required to ensure that cookie functions do not change and to periodically sample the model’s results for accuracy. A team will not only improve the model’s performance by validating results which then can be fed to the model, but also to ensure that cookie functions and descriptions that are fed to the model are accurate and up to date.