Rodrigo Rodriguez

07/05/2024

Applied Data Science

DSC680 (SUMMER)

**Project Whitepaper Draft:**

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

**Business Problem**

With the recent laws passed regarding consumer data privacy, such as CCPA, our legal and cyber-security team now heavily focuses on identifying and correctly categorizing online cookies. This project focuses on automating cookie classification by leveraging the use of an NLP classification model.

**Background and History**

WM (Waste Management) is the largest waste and recycling services provider in the nation. The company is focused on growing the online sales channel. Doing so requires staying compliant with online privacy and security laws. Recent data privacy laws state that website users must be given the choice to opt-out of any online cookies not categorized as (Essential). Manually categorizing cookies is a time-intensive process and will require many months of work for a team of analyst to perform. Leveraging a ML model will speed up the process of categorizing these cookies and in doing so, will mitigate WM’s exposure to legal risk.

**Data Explanation**

The dataset contains several categorical variables with the target variable being “CategoryID”, this feature indicates what category the cookie will be placed in. Most variables do not add predictive value to the model and will have to be removed during data preparation. The dataset was collected by an analyst and enables supervised learning. Once the non-predictor variables are removed, we are left with only three variables: “CategoryID”, “C\_Source”, and “Description”.

**Methods**

During EDA the distribution of different key variables were plotted. The primary finding was the presence of class imbalances in the “Category” column as seen in Figure (1). To ensure adequate representation in training and test sets, stratified sampling was implemented. After EDA the data must be prepped before processing. This includes dropping any non-predictor variables to remain with the three key variables listed above under Data Explanation. Then dummy variables will need to be created from the “C\_Source” categorical variable. Lastly, the natural text data in the “Description” variable will need to be processed and the dataset will have to undergo text vectorization.

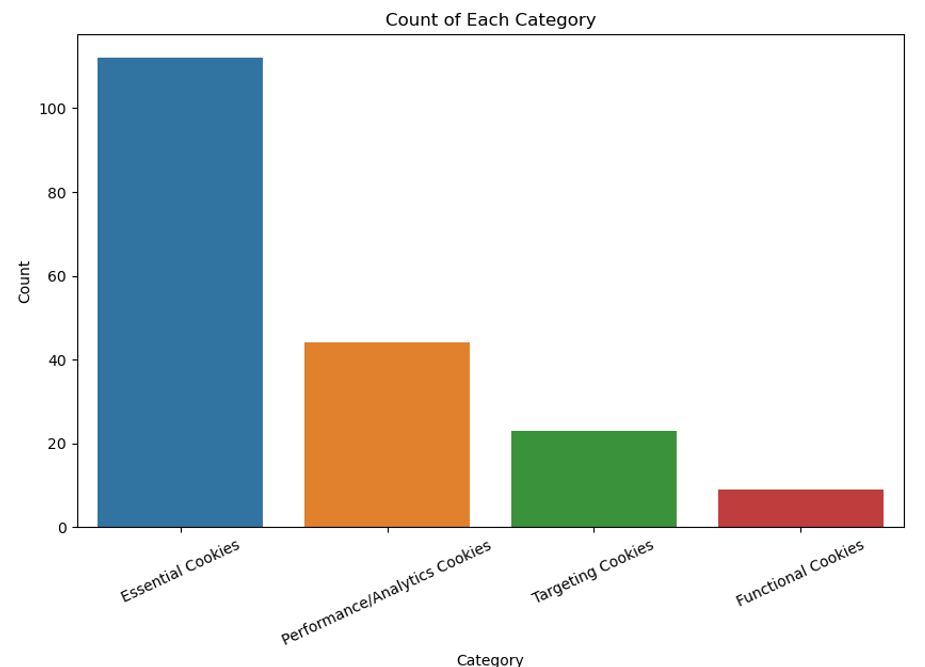


Figure 1

**Analysis**

Three classification models were tested for this project, Random Forest, Logistic Regression, and Naïve Bayes. After training and testing each model a classification report and confusion matrix were used to evaluate performance. The classification reports and confusion matrix were very similar for all models. Figure (2) shows a report and confusion matrix for the Logistic Regression. The class imbalances can be seen in the low f1-score for class 2 and class 4.

A screenshot of a computer screen

Description automatically generated

Figure 2

**Conclusion**

All models exhibited the same relationship with macro F1 and weighted F1. The macro F1 was lower at about .72 and the weighted F1 was higher at about .84. The performance drastically drops for the minority classes (Class 2, and 4). The results of this project look promising. An accuracy of above 85% and high F1 scores for classes with adequate representation in the data prove that the model is capable of correctly classifying an online cookie based on its description. However, no model should be deployed as is. Performance can be improved by addressing class imbalance.

**Assumptions**

We assume that the descriptions for each cookie are adequate and true. Initially, the descriptions were collected manually by researching the function and use of each cookie. Moving forward, a cookie database API will be used to retrieve cookie descriptions based on cookie names.

**Data limitations**

This data is limited in the number of records available for training and testing. A manual effort was made to collect the supervised learning dataset used for this project. However, the model could improve performance by having access to a more robust dataset; preferably one not affected by class imbalances.

**Challenges**

The main challenges for this project were the limited dataset, class imbalances, and short project time period. As listed under Data Limitations, the class imbalances negatively impacted the model’s performance. Additionally, if there was more time to work on the model hyper parameter tuning could be done to further improve model performance.

**Future Uses and Recommendations**

This model can be scaled and used for all web cookies across all of WM owned domains. Alternatively, this project can be leveraged by other companies looking to classify their online web cookies. My recommendation is to move forward with optimizing the Logistic Regression model over Naïve Bayes and Random Forest. The Logistic Regression model outperformed Random Forest, and can be further improved by tuning hyper parameters, while the Naïve Bayes is limited in this aspect (Yiu 2019).

**Implementation Plan**

After addressing the challenges and implementing recommendations, the model can be implemented by feeding new data from the cookie database API. The data will be fed to the model for classification and a random sampling can be taken from the model’s results to validate manually. If manual validation is not passed, then the model will have to be further optimized.

**Ethical Assessment**

We are relying on the cookie descriptions provided by API to be accurate and true to properly assign cookie categories. Continuous cookie management will have to take place to monitor and ensure that cookie functions are not changed by the cookie hosts and that cookies are correctly classified.

**Potential Audience Questions**

1. What cookies were affected by class imbalances?
2. Why are certain cookies underrepresented in the dataset?
3. How was the data split between the training and test sets?
4. If all models performed about the same, why is Logistic Regression recommended over the other evaluated models?
5. Why were so many variables dropped during data preparation?
6. Why is F1-Score the primary metric to focus on in the Classification Report?
7. Why did text need to be processed before modeling?
8. Did exploratory data analysis reveal any other insightful information?
9. Would this model be required after all cookies are categorized?
10. Do we still need a cookie management team after implementing this model?

**Sources/Appendix**

Yiu, T. (2019, October 24). *Understanding The Naive Bayes Classifier*. Towards Data Science. Retrieved May 24, 2024, from https://towardsdatascience.com/understanding-the-naive-bayes-classifier

Pramoditha, R. (2023, November 29). *Logistic Regression for Multiclass Classification — 3 Strategies You Need to Know*. Medium. Retrieved May 24, 2024, from https://rukshanpramoditha.medium.com/logistic-regression-for-multiclass-classification-3-strategies-you-need-to-know-0a3e74574b96

Komnenic, M. (2022, January 14). *The Different Types of Internet Cookies Explained*. Termly.io. Retrieved May 24, 2024, from https://termly.io/resources/articles/types-of-internet-cookies/