Springboard Data Science Capstone

Capstone Project #3

Sentiment Analysis

Abdul Hannan

10/2022

1 – Introduction

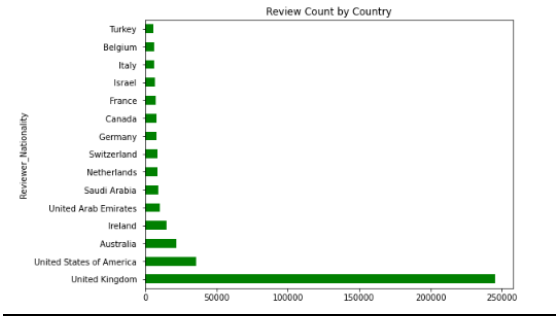
The purpose of this project is to extract emotions and give them a rating of positive or negative given textual data. We will be using a technique from Natural Language Processing (NLP) called sentiment analysis. It is an important step to gauge customer response, the data collected from customers using multiple portals, and based on taking these steps, we can target customers differently or change our target audience.

Along the process we learned that a lot of the variables were generic, and that we would not need them while we are building the model as it creates extra noise and room for error. Our managers are interested in a model that does not target a review as false negative, we do not want an error of a model affect any hotels business, so the method we will focus on will be the Recall method. In summary, we used the textual data and the score that the reviewer assigned and used that to create a model, to get there we cleaned the textual data using necessary steps that will be discussed later. Our best model was Naïve Bayes which received a recall score of 0.9.

2.1 – Data Acquisition and Wrangling

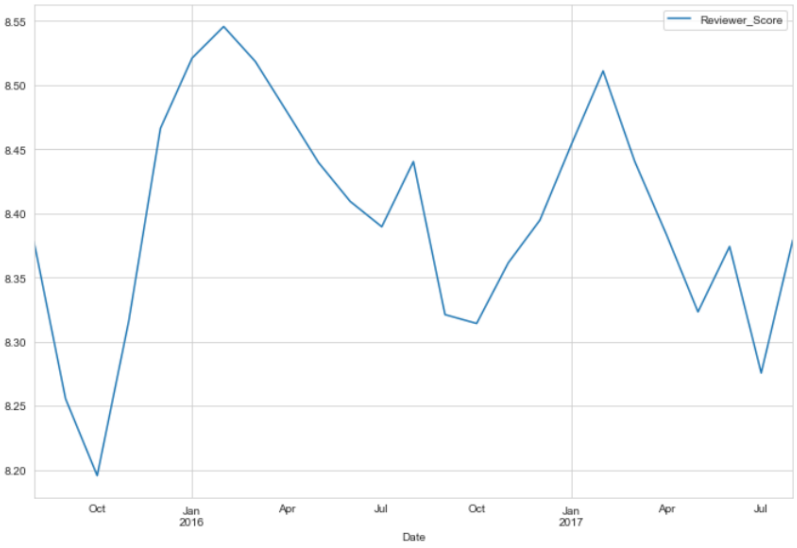
The total number of variables we have are 17, we will use them to create a story for our Exploratory Data Analysis (EDA). We saw that our reviewer nationality column was heavily skewed as almost half of the reviewers were from United Kingdom. We also see that all the hotels are consisted in the following six countries, United Kingdom, Spain, France, Netherlands, Austria, Italy, all are within the European borders. After finding out that there are no missing values in the interested variables (Positive, Negative reviews, and Reviewer Score), we are ready to move on to the Exploratory Data Analysis.

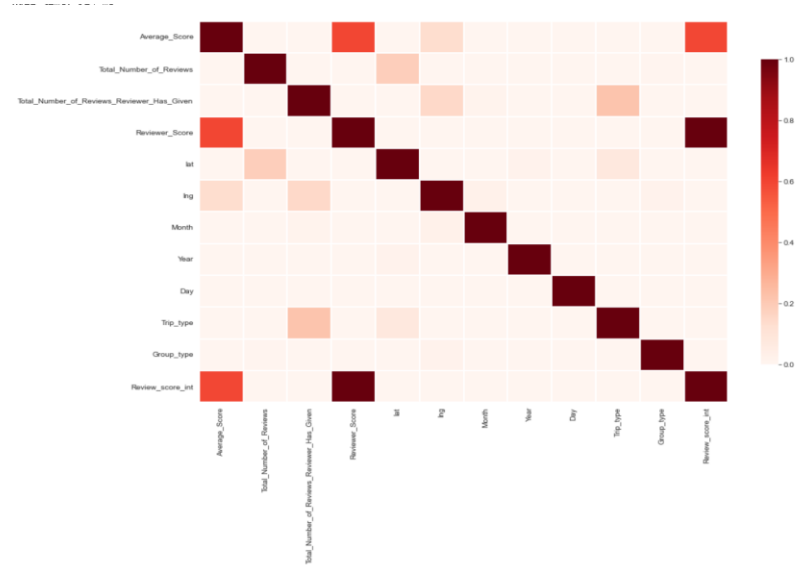
2.2 – Storytelling and Inferential Statistics



We learn a couple things from EDA, some of which are most of the hotels were based in France out of the six countries, and for the majority part, the reviews came from UK and USA. Another thing that struck out was that countries whose first language is English tend to give a higher rating than countries who do not have English. United States, Australia, UK, Canada are amongst the top 5 nationalities who give a higher score, whereas UAE, Saudi Arabia, Turkey are amongst the last 3. Among the 1494 that received a review, France had the highest number of hotels with 458, followed by United Kingdom with 400. On top of that, only 100k+ reviews came from people who went to the trip solo, mostly for business purposes, while the rest came with a group of people with 2 or more. We proceeded to look further into, at what time of the year were

customer more satisfied. Starting from Oct – Jan, there is an increase in scores and it happened both years. Moving along, we plotted a heatmap to gain further insights between features.





2.3 Baseline Modeling

This is where the fun starts, to generate a Machine Learning model, we need numbers and not text, so we need to convert text to numbers. Before the conversion, we need to clean the text. Selecting three columns (Positive_Review, Negative_Review, Reviewer_Score), we then plan

our next steps to clean our textual data. For our numerical variable Reviewer_Score, a class 1 will be assigned if the score is between 7.5 - 10, and class 0 if the score if the score is between 0 – 7.5, here our target class is 0, because we have more positive reviews (78%), compared to class 0 (22%). To get our textual data ready for the model, we then lowered the text, remove punctuation, removed stop words, lemmatized the text (for relevant results), converted no reviews to empty spaces. After creating our train test split with a respect ration of seventy-five and twenty-five percent, we were ready to convert our textual data into numbers. We used three vectorizations technique to do that, Count Vectorizer, Term Frequency Inverse Document Frequency Vectorizer, and Hash Vectorizer. Creating a generic Logistic Regression model with all three vectors produced almost the same models. We also learned that the model is overfitting, having model perform well on training data and not on our evaluation set. Again, we are aiming to get a high recall on our target class, as we want to be sure that our model does not create false negative cases, which will affect a hotels operation and business.

2.4 – Extended Modeling

Now that our textual data is cleaned, and our target feature is set, and our target class is identified, we applied resampling methods to our skewed data to even the ratios of both classes. Two techniques were applied, Random Under Sampler (RUS – Under sampling method), and SMOTE (Over-sampling method). We then created four models each time, one for every resampling technique and vectorizers combination, and three different models in total XGBoost, Naïve Bayes, Random Forest. Our preferred method is optimizing recall because we want to avoid the case of false negatives, as it will affect a hotels rating, which in turn will be bad business based on a machine error, that is labeling a hotel rating as negative, when it is positive.

3 – Findings

	Model	Sampling	Vectorizer	0: Precision	0: Recall	1: Precision	1: Recall
2	Naive Bayes	SMOTE	TF-IDF	0.50	0.85	0.95	0.77
3	Naive Bayes	RUS	TF-IDF	0.50	0.85	0.95	0.77
0	Naive Bayes	RUS	CV	0.50	0.84	0.95	0.77
1	Naive Bayes	SMOTE	CV	0.52	0.82	0.94	0.79
6	XGBoost	RUS	TF-IDF	0.62	0.55	0.88	0.91
7	XGBoost	SMOTE	TF-IDF	0.62	0.55	0.88	0.91
4	XGBoost	RUS	CV	0.51	0.21	0.81	0.95
5	XGBoost	SMOTE	CV	0.51	0.21	0.81	0.95
8	Random Forest	RUS	CV	0.49	0.09	0.79	0.91

This is where we as data scientists make our decision, and that is deciding between a trade-off, where recall does better than precision, and that is present in our best model. Naïve Bayes model (which are more efficient) does well in our scenario rather than our decision tree-based model. Naïve Bayes model with TF-IDF regardless of our sampling technique does better, as we achieved a recall of 0.85, but we do have a low precision of 0.5. We came a long way from our generic model which was built on a not-cleaned textual data, and was overfitting. Achieving a recall of 0.85 ensures us that it will be able to identify the cases of false negative 85 times out of a 100. There will be instances where a hotel will get a bad rating based on a models error, but having a recall of 0.85 for our target class is a step towards the right direction.

4 - Conclusion

After reviewing the context of our dataset, we used all variables but 3 to conduct insights into our study and conduct an EDA analysis. All the hotels were based in 6 European countries with France having the most hotels. We found the reviewers giving more scores towards the end of the year, as that can mean the season of festivities, which means a reason to travel. We had reviewers from many Nationalities however nationalities that had English as their first language

tend to give higher scores than ones who don't. Some of the words that were most common while creating the vectors were (minibar, fruit, basket, clean, sheet, etc.). After creating our two vectors (2 possible models) we used two types of resampling techniques (2 more possible models) and three types of Machine Learning models (2 X 2 X 3 = 12), we would be expecting 12 models. In our notebook, we have 9 and found that tree-based models tend to be doing worse, and are very time costly, instead of our Naïve Bayes model. Naïve Bayes with term frequency – inverse document frequency vectorization (with both sampling techniques) produced us a recall of 0.85, the highest among all, ensuring that the false negative cases is something that the model would do better at to avoid an external cost to a hotel's reputation.

5 – Recommendations for the clients

- Given that our textual needed cleaning, and let us be honest, it is not a big dataset considering we are working with reviews and textual data, we can apply a few techniques. Data augmentation to reproduce our data which in turn might lead to better accuracy for our models, we will also be able to make more advanced models and reduce bias in our models.
- Second approach can be Keyword Extraction, as it will help us find meaningful insights in a short span of time and help us identify the topic / problems customers have based on their reviews.
- Text Summarization is another approach that can save us time, in another words, paraphrasing of our text and keeping useful words without having to read for word to word.