Milestone 3: White Paper

Felipe Rodriguez

Bellevue University

DSC 680 Applied Data Science

Professor Amirfarrokh Iranitalab

May 31, 2024

Project Overview/Background

In the city of Boston, Airbnb rentals continue to grow. Airbnb is looking to analyze and understand sentiments of reviews. The sentiment analysis project is meant to categorize reviews into positive, negative, and neutral reviews. Additionally, by identifying the sentiment of the review, a model can be created to predict the sentiment. The fields used to make predictions will be comments and sentiment, which will be created in the project.

Data Overview

The data is being obtained from Kaggle and is sourced from the Airbnb. The data contains the following information: index, listing_id, id, date, reviewer_id, reviewer_name, and commented. There will be a field created named sentiment that will be generated during the project. The data will allow for the analysis of the comments and create sentiment. The fields in the data contain the following information:

Data Dictionary

column	description	data_type
index	index	int64
listing_id	Identifier for listing	int64
id	Identifier for review	int64
date	Date of review	object
reviewer_id	Identifier for reviewer	int64
reviewer_name	Reviewer Name	object
comments	Comments made by the reviewer about the listing	object

The data preparation phase requires cleansing and prepping the data so that it is fit for use. The first portion will involve cleansing the comments for sentiment determination. This will involve removing stop words and punctuations. Once the comments have been cleaned, they will be used to create the sentiment column.

Methods

The methods used in this project will include various visualizations of comments, sentiment, and results, as well as a Naïve Bayes model to predict sentiment. The visualizations will include distribution of the sentiment field, word cloud, and ROC curve of the results. The world cloud created will highlight the most common words used in the reviews and the ROC curve will provide insights into the model performance.

Since this model gave high results, the same model was conducted for the precipitation and snow depth variable. The model results for precipitation and snow depth did not perform as expected. While the same type of model was created, these two provided lower R-square scores. Because of this a Seasonal Autoregressive Integrated Moving Average or SARIMA model was used instead (see appendix B). This additional model on the other two variables provided low accuracy as well.

The model chosen for this project was a Naïve Bayes classifier. The Naïve Bayes classifier is used because it is "simple and computationally efficient, performs well with small datasets, and is easy to interpret" (Chaudhuri, 2024). This model created uses the reviews and sentiment field to generate prediction. This model performed very well with a high R-square score (see Appendix A). R-square is used as the metric for determining model accuracy because R-square shows how well the data fits the regression model (Taylor, 2023). Additionally, F-1 score can be used to determine accuracy as well. F-1 score is "evaluation metric that measures a model's accuracy and combines the precision and recall scores of a model" (Kundu, 2022). The weighted F-1 score also performed very well. The weighted score is a combination of the different classes. Class 0 and 1 performed lower than Class 2, which was positive reviews.

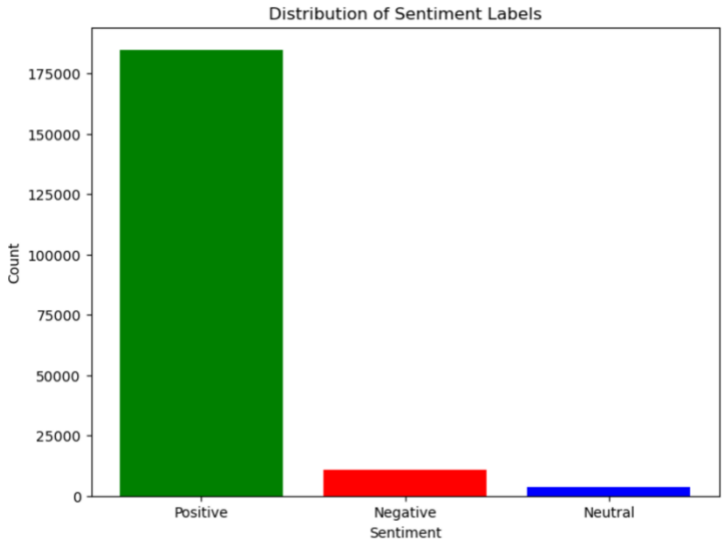
Assumptions

Prior to analyzing the data, the following assumptions are made:

- The data provided and gathered is used with the consent of reviewers.
- The model assumes the data is balanced and contains a variety of reviews.

Analysis

The data explored gives insights on the reviews in the Boston Airbnb market. The first plot gives insight on how the reviews are distributed. The graph on the right shows that there are more positive reviews than negative or neutral. The next visual is a word cloud. The word cloud uses the individual elements of the comments and displays their size based on occurrence. When viewing the word cloud, a lot of the words are positive in nature. This could be due to the number of positive reviews versus negative and neutral. One of the items in the world cloud that might be worth noting the frequency of 'walking distance'. This comment might be perceived as neutral but can be considered a positive for some users. When looking further at the comments, there are some words that are not relevant such as 'ca' and 'nt', some of this has to due with the preprocessing of the data. This does not affect the model.



Challenges/Limitations/Ethical Considerations

The challenges in this data were found when creating the model. One of the first items is to successfully remove stop words and punctuations. Without this, the model would not be able to perform appropriately. Another item is that the comments need to be vectorized prior to being added to the model. The exploration of the model found 100% accuracy which was due to using fields derived from the sentiment field. This was changed and comments and sentiment were used for the model.

One limitation of this model is that the model is based on the reviews available. This data has a high concentration of positive reviews, which is shown in the graph as well as the F-1 scores (Appendix A). When it comes to Naïve Bayes, there could be oversimplification of reviews. Which means that the model could have limited context when creating predictions.

An ethical consideration with this model is to ensure that the reviewer's privacy is protected. Another item to consider is the consent of using reviews and the right to control the user's data and information.

Future Uses/Recommendations/Implementation Plan/Conclusion

The Naïve Bayes model performed well at predicting sentiment with a high R-Squared score and F-1 score. It is recommended that this model be implemented for use. The best way to implement this model is to develop a web application that will allow a user or agent to input a review and retrieve the sentiment based on the model that was developed. An example of a program that will allow for this ability is Streamlit. To use this, a model needs to be created, a connection or method to access the model, and lastly the Streamlit server (Shiledarbaxi, 2021). This method does not need knowledge of web application programming languages.

In this project, reviews were analyzed to understand the sentiment and use the sentiments to create a model to predict sentiments. The model performed well with a high R-squared score and F-1 score. This analysis uncovered that the data contains many positive reviews. To improve the model in the future, it is recommended to include more variety of reviews.

Appendix A

This appendix includes the results and findings of the Naïve Bayes model created. The model has comment as the feature and sentiment as the target. The results model resulted in a R-Squared of 0.95. To ensure model accuracy, there is a classification report created that displays F-1 Score, Recall, and precision. Below are the results of the classification report.

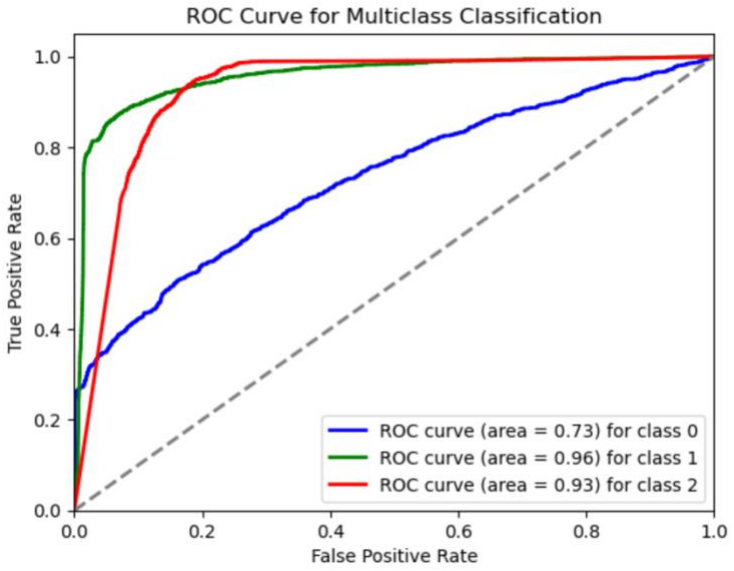
```
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.25      0.37       763
           1       0.71      0.63      0.67      2127
           2       0.97      0.99      0.98     36932

    accuracy                           0.96     39822
   macro avg       0.79      0.62      0.67     39822
weighted avg       0.95      0.96      0.95     39822
```

The results of the classification report show that classifier 2 has very good precision, recall, and F-1 score. This could be because there are more positive reviews in the data set.

Additionally, an ROC curve is created to gain insight on the results. Below is the ROC curve. The ROC curve shows the performance of the classifier in distinguishing one class from the other. This curve shows that the classifier performs very well for class 1 and class 2. Improving Class 0 would result in better model performance.



Reference:

Chaudhuri, K. D. (2024, May 7). *Building naive Bayes classifier from scratch to perform sentiment analysis*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/#:~:text=Advantages%20of%20Naive%20Bayes%20for,results%20based%20on%20word%20probabilities.

Kundu, R. (2022, December 16). *F1 score in Machine Learning: Intro & Calculation*. V7. https://www.v7labs.com/blog/f1-score-guide#:~:text=for%20Machine%20Learning-, What%20is%20F1%20score%3F,prediction%20across%20the%20entire%20dataset. https://analyticsindiamag.com/how-to-deploy-time-series-forecasting-models-using-streamlit/

Shiledarbaxi, N. (2021, March 8). *How to deploy time series forecasting models using StreamLit*. AIM. https://analyticsindiamag.com/how-to-deploy-time-series-forecasting-models-using-streamlit/

Taylor, S. (2023, November 22). *R-squared*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/r-squared/