Model Details

- **Created by:** Minh Vuong Tran, Gabe Palomino, and Allie McBride (12/11/2023)
- **Purpose of the model:** To determine an accurate genre-labeling of a book based on its summary and title.
- **Statement about ML technique(s):** The primary ML technique used in our model is logistic regression, which is well-suited for classification-based models (as demonstrated in notebook 5.1). Moreover, the model utilizes a TF IDF-scoring process, which we found conducive towards the naturally language-centric analysis needed for the model.

  Utilizing a WEB-Scraped data set published on the data science platform Kaggle, the model is trained on 4,567 entries containing the title, summary, and genre of various novels. Its primary capability is the labeling of novels based on relevant input text (namely the aforementioned titles and summaries). Furthermore, its labeling is limited exclusively to 10 genre categories: Thriller, Fantasy, Science, History, Horror, Crime, Romance, Psychology, Sports, and Travel.

Intended Use

- **Who/What use cases were considered:** Due to its relatively small labeling set, the model's intended use is primarily to streamline or assist certain processes. For example, librarians could use the model to provide initial labels for a large collection of books before re-labelling them with more accurate/specific classifications (if needed). Additionally, online retailers or content recommendation systems could couple our model with other AI implementations to enhance suggestions or generate reviews - notably as part of a larger process.

- **User spectrum:** Although the use cases above are examples of intended applications, the model is NOT intended for more sophisticated deployments such as in literary or scholarly work, as the generated labels are not intended to be wholly representative. Rather, they are intended to provide a possible genre that the novel *may* be classified as.

Factors

- **Groups:** Our Data was organized by title, genre, and summary.
- **Instrumentation:** Windows 11 64 bit
- **Environment:** Our model operates on Visual Studio Code using Python and Python installations such as Sci-kit Learn
- **Relevant Factors:** Title, Summary, Genre (Possibly Title and Summary length)
- **Underlying:** Year, Author, Publisher, Edition, Language (Not included in our data file)
- **Evaluation Factors:** Title, Summary

Metrics

- **Performance measures:** *Note: Although our group created 3 variations of our model, only the primary model (which accepts both text titles and summaries) will be discussed*. The genre-labeling model has an overall accuracy score of 0.903. In the context of our specific model, this means that a correct labeling is assigned within the top 3 predictions of our model 90.3% of the time. Furthermore, the mean reciprocal rank is 0.7704 (reference the quantitative analysis section for a specific breakdown).

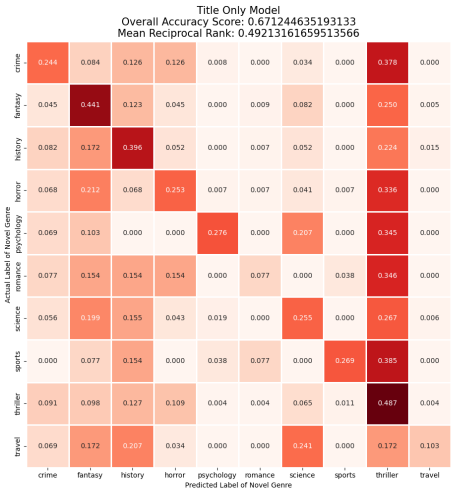
Evaluation Data

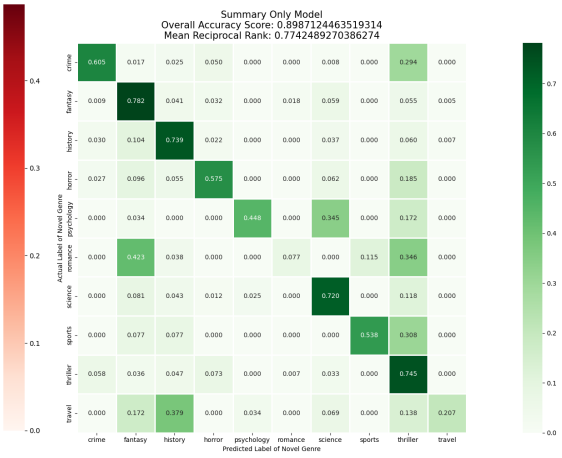
- **Data overview (ref. datasheet):** The original data used for training was created by Arthava Inamdar (see Datasheet). It was created using goodreads data processed through Selenium. Top books from the selected genre pages were extracted and compiled into a singular dataset. Books were sorted into columns of genre, and were accompanied by corresponding title, summary, and index columns. Notably, while each entry was unique, certain ones contained identical titles or summaries.
- **Processing work:** To process the data, we performed various EDA work. During this process, we found no missing, incomplete, or corrupt data; however, several books and summaries were duplicates assigned with different genres. After consideration, we elected to keep such duplicates within our training data to hopefully expand the accuracy of our model in real-world contexts and to introduce degrees of nuance when attempting to classify novels.

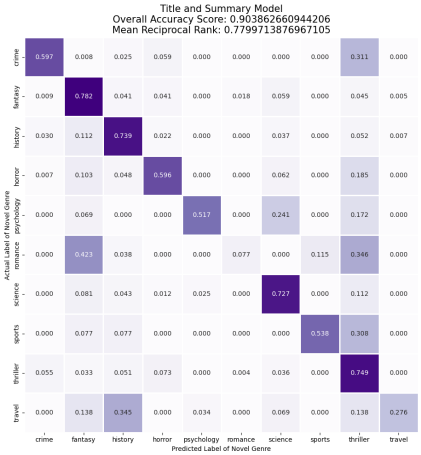
Training Data

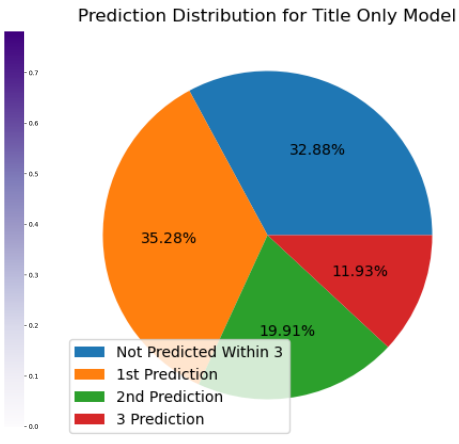
- The training data is an extraction of the evaluation data using a concatenation of title and summary information.

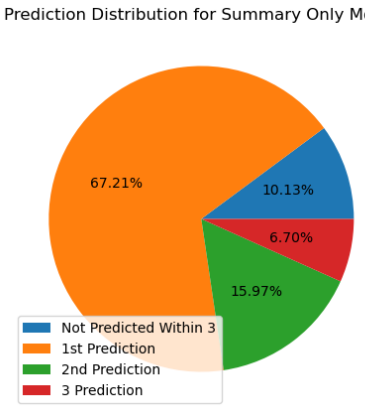
Quantitative Analysis

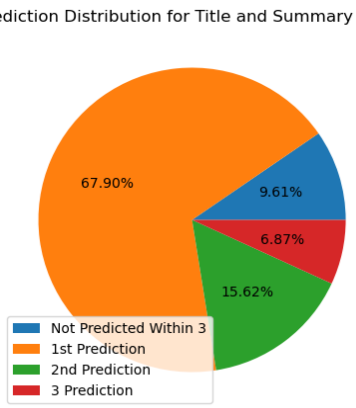


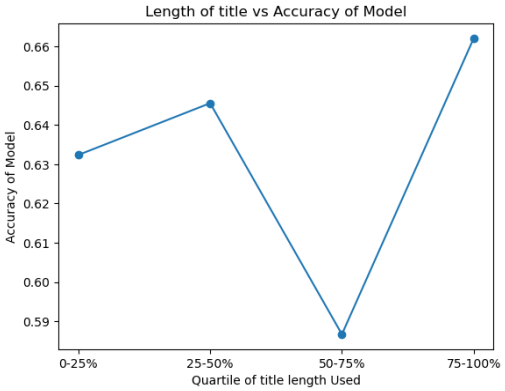


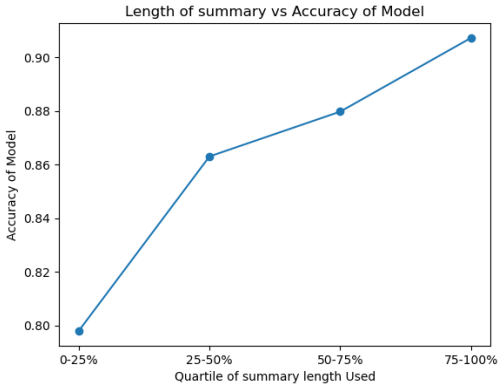


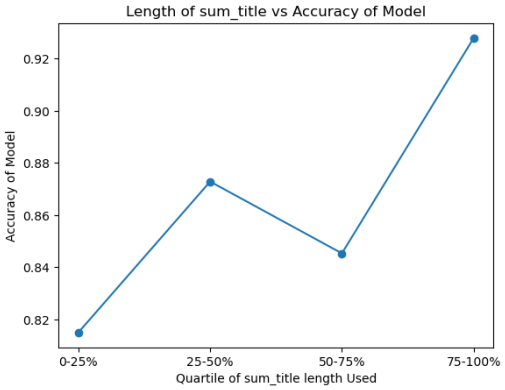












Ethical Considerations

- As highlighted in Keven Guyan's *Queer Data* it is important to understand the people your data is coming from, and who it will affect. To appease this ethical concern, we made sure our data was coming from sources we could fully understand, (books in English, following western genre categories). We also ensured that our model would have a positive effect on this group: making it easier for them to search their favorite genres. However, this lack of book diversity makes our data cater less to social justice. As expressed in N. N. Jones and R Walton's *Social Justice*, social justice puts great importance on diversity and equal opportunity for all. Because our model performs solely for western genre books in English, it is not very diverse and socially just.
- The use of a Web-Scraped data set obviously poses a consideration of ethical expediency. However, we justify its use through the minimal importance it plays within

our model. The Web-Scraping process simply allows for an easier transfer and compilation of similarly grouped data. However, the quality assurance, exploratory work, and critical analysis of the data was still maintained. A more professionally collected and tailored dataset - including other relevant information - would nevertheless have been more ideal.

Caveats and Recommendations

- In our EDA work, we stumbled upon data with repeated titles and summaries yet different genre labellings. We ultimately decided to keep them, as they illustrated a level of nuance in genre classification that we found beneficial to our model, i.e. a novel typically has multiple valid classifications and should be trained as such.
- The model has a slight bias towards data with a larger occurring frequency in the training data, namely the genres of thriller, fantasy, science, and history. However, unrelated “pockets” of significant confusion rates also exist.
- Since the model deals strictly with a limited number of genres, we strongly urge that applications of this model be centered primarily within this range or adjacent to it. For example, the genres of history and science can be extended to both fiction and non-fiction texts. However, novels such as autobiographical, cooking, etc. should not be used in conjunction with the model.