BRAINSTATION: DATA SCIENCE

CAPSTONE PROJECT

Collaborative Filtering and Content Based Recommender System

Author: Uldis Knox

Date: December 11, 2022

Business Question: Could you effectively recommend to someone a new beer or beers to try, using only their existing reviews of beers they have tried and reviewed previously.

Background: Data is very powerful, you can derive significant amounts of insight from small amounts of data. In this specific instance, we can look at a User history and see what they have liked and disliked, similar to the recommender systems used by Amazon, Netflix, and online retailers. This recommender system is similar in that it uses historical data to recommend new, similar items that a User has not rated to what a User has rated.

This recommender system adds value in that it is location agnostic, brewery agnostic and the dataset is international. Previous iterations of recommender systems for beer rely on regionality, so only capturing certain countries in their process, or they are brewery specific, so we are looking at a small sample of beer types, from a very specific place.

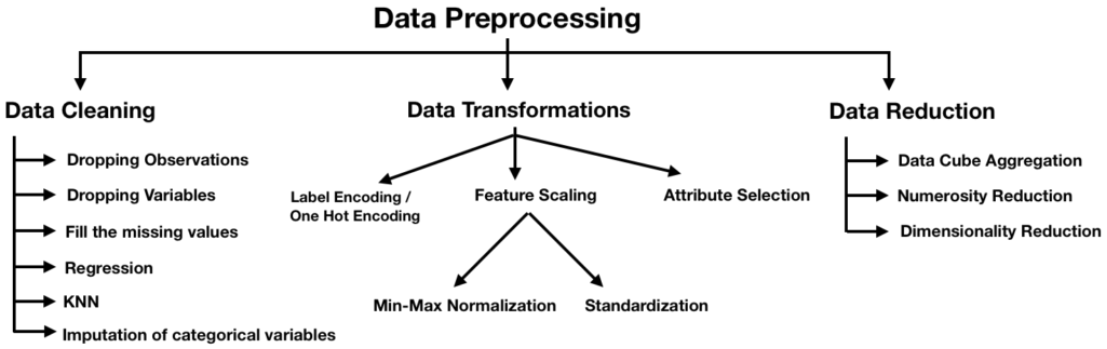
Dataset: Our data was acquired from this web location: https://data.world/socialmediadata/beeradvocate.

This dataset was originally web-scraped from the website Beeradvocate by computer scientists at Stanford who were working on projects involving online reviews.

- J. McAuley, J. Leskovec, and D. Jurafsky. Learning attitudes and attributes from multi-aspect reviews. ICDM, 2012.
- J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013.

The dataset consists of beer reviews from beeradvocate. The data span a period of more than 10 years, including all ~1.6 million reviews up to November 2011. Each review includes ratings in terms of five "aspects": appearance, aroma, palate, taste, and overall. Reviews include product and user information, followed by each of these five ratings.

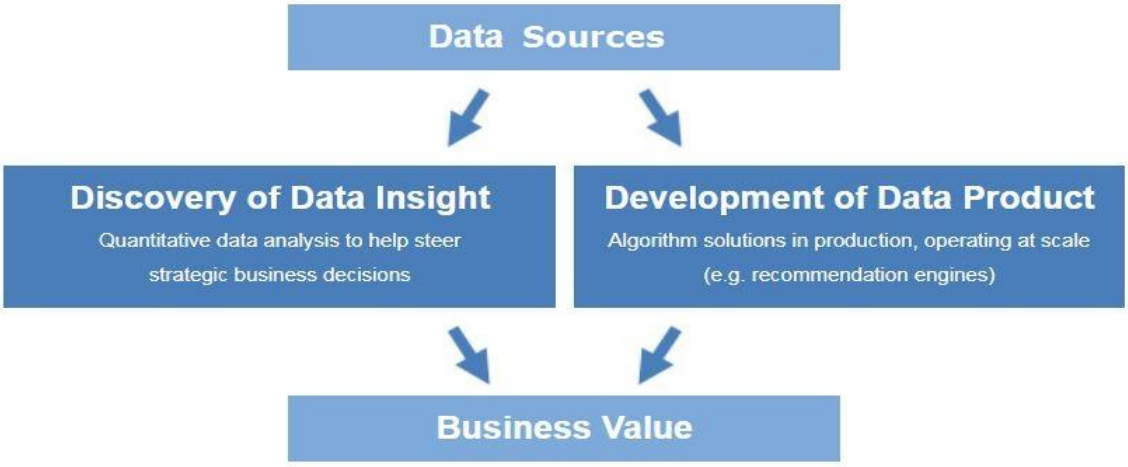
Data Preprocessing



The data cleaning and preprocessing began with loading our data into Jupyter Notebooks and looking at what it contained in terms of features (columns) and data points. Once we had taken a good look through our data and determined which features to keep and which to drop, we removed those which features we didn't need, and checked for missing values in what was kept. Some of the missing data was dropped from the dataframe, others were populated with the mean value of the feature.

We did add three (3) features to our dataframe, those being City, Country and Taste. We had to manually collect the City and Country data for each brewery as the scraper script we ran encountered issues with data missing from the site as well as language barriers from accents above consonants and vowels to letters that do not exist in the English language. Taste data was also manually collected from a separate website and both the location (City, Country) and Taste data were merged into our original dataframe.

Insights, Modelling and Results



The final dataframe was inspected for distribution of our rating features, we also looked at how many countries were represented in the data (Tableau map) and to what frequency the countries were represented. Next we looked at the top rated beers in the data by review and by frequency of review and finally the top beers by style.

We proceeded to set up a k-Nearest Neighbors (kNN) model, which was run on our data by converting the 'review_overall' feature to be a categorical variable. We applied two scalers to our data after the initial kNN model, followed by finding the best value for k and hyperparameter tuning before fitting our optimal parameters to the kNN and running one final time. Our models all performed very similarly, given so few features being applied.

Logistic Regression followed, which was paired with a SMOTE analysis to balance our data, we did not see a lot of variance in our models again, similar to kNN.

Finally, we developed the Collaborative Filtering and Content Based Recommender System. This is our pièces de résistance. The Recommender takes inputs from the data in the form of individual user reviews, analyzes them and predicts beers the User has not yet rated, but are similar to ones they have rated and recommends them, along with predicting the User's mean rating for those recommended beers.

Findings and conclusions

When the project started, the intention was to have a dataset that included text-based user reviews, there just wasn't a dataset large enough for the project that included them, so we pivoted to using the numeric rating features. The added features of City, Country and Taste did not influence the project much besides some visuals and outputs and I would not go through the agony of manual data collection for them again.

Upon completion, the Recommender System that was built is aligned with initial project outlook and expectations, even though it was not planned to be a Collaborative Filtering, Content Based Recommender.

I have spoken with several people in the restaurant and brewing industry and there is some interest in a product of this nature. I would like to find a larger, more recent dataset that includes User reviews to improve upon the models, add some complexity, and improve the Recommender System output.