Gender Classification for Perfumes

New Sense Company wishes to create bestseller perfumes and market it to the right crowd at the right time. Before launching new perfumes, they would like to identify **the gender to market it to**. A perfume is made up of notes(distinct oils) which in turn make up perfume accords (an oil blend). The fragrance of a perfume is defined by its notes and whether it is part of the Top/Middle/Base layer of a perfume. Given the notes of a perfume, we would like to classify it as unisex/men/women's perfume.

1. Data

fragrantica.com is one the world's largest international perfume reviewing websites. Here users can login and rate perfumes. Details about perfumes, their notes, along with user ratings are available here.

With over 50,000 perfumes and their ratings, this Kaggle webscraping project is a sufficient size to develop a good classifier model. To view the fragrantica.com website, the original Kaggle datasets created by David Cohen, and the profile report click on the links below:

Fragrantica website

Kaggle Dataset

Profile Report

Figure 1: *An example of input data*

2. Method

This is a multi-class classification problem. We are trying to classify a perfume as suitable for unisex(2)/ men(1)/ women(0) based on the notes of a perfume.

Scikit-learn supports many Classifier algorithms.

1. Naive Bayes
2. Random Forest
3. Adaboost
4. KNN Neighbor
5. Logistic Regression
6. SVC

We will attempt all of these to find out which classifier achieves the greatest accuracy in this case.

3. Data Cleaning

Data cleaning notebook.

In this classification problem, we need to mainly only look at the notes that make up the different layers of a perfume. So all other fields that were unnecessary were removed.

a. **Problem 1**: This is a non-labelled dataset. So the first priority is to create a label. We have user-reviews for each gender pool which were used to classify a perfume as either unisex/men/women

b. **Problem 2**: The *gender* details we received in our dataset does not match with the gender that we can extract from the *title* of a perfume. Hence the original *gender* columns were dropped and a new *gender* was extracted from the *title* column.

c. **Problem 3**: The notes of a perfume are given in 20 fields with a prefix of *Top/Middle/Base* based on which layer they appear in the perfume. This data is not in a usable format. Hence notes in each layer were collated into *top_notes*, *middle_notes* and *base_notes* columns. We also created a *notes* column to keep all the notes of a perfume in a single column.

d. **Problem 4**:The *notes* values are all categorical values. We cannot pass them as such to an ML model. We have a total of 1440 distinct notes in our perfume dataset. That is too much information. Hence I decided to create columns for all distinct notes that appear in 100 or more perfumes. This gave us a total of 299 popular notes. We also wished to preserve the layer information of a note.

   So if a note appears in Top Layer - its value is 2^2, Middle layer - value is 3^2, Base layer - value is 4^2. If a note appears in more than one layer, the sum of the values are taken. Ie. a note appearing in top and middle will have a value of $2^2 + 3^2$.

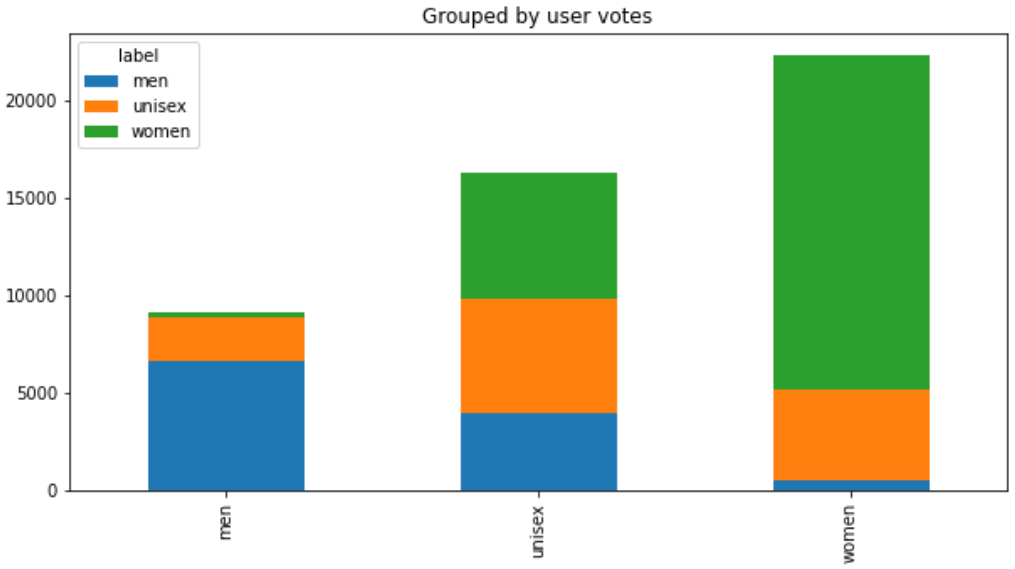
4. EDA

EDA Report

The perfume industry mainly caters to the female population, as can be seen in the graph below, with the female perfumes ranging close to 23,000 in number. But **20%** of these perfumes have been voted as suitable for **unisex**, rather than just for females.

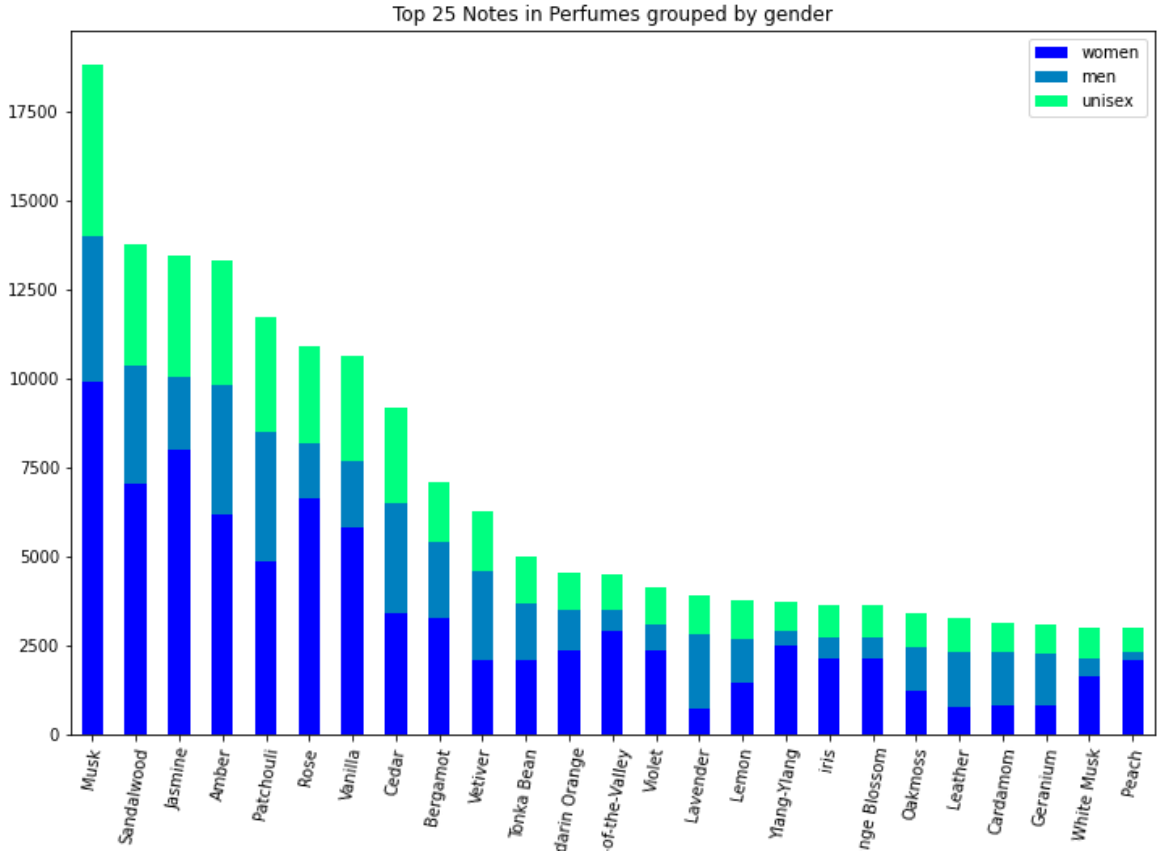
The perfumes originally marketed as suitable for **unisex**, are not perceived as such by the customers. **40%** of these perfumes have been voted as suitable for *women*, **36%** as suitable for *unisex* and **24%** as suitable for *men*. The opinion seems to be divided.

The *men* category has the least number of perfumes. But **24%** of *men's* perfumes are considered as *unisex* by customers.



We also attempted to see if there are any marked preferences for specific perfume notes by a particular gender. There are some notes, like 'Musk' which is present more in Women's perfumes, while we also have some notes like 'Cardamom' that appear more in men's perfumes.

So there is a clear preference for specific notes by a gender.



5. Algorithms and Machine Learning

ML Notebook

I chose to work with the Python scikit-learn for training my classification system. I tested the dataset on the following algorithms:

Classifier Name	Accuracy	F1-Score	Precision	Recall
Naive Bayes	**58**	**51**	**51**	**52**
KNN	52	45	46	45
Random Forest	58	50	51	52
Adaboost	59	47	51	50
Logistic Regression	59	47	52	50
SVC	60	46	53	50

If we go by only the accuracy, the best model to choose will be the SVC Model. But all the scores appear good for Random Forest and Naive Bayes models.

The dataset does not capture the relationship between the independent features and the dependent feature correctly. The models have failed in classifying them as per the business requirement.

6. Predictions

As the models have very low accuracy, we did not make it into the prediction phase.

7. Future Improvements

NA

8. Credits

My mentor, Mr. Ankit Gupta, played a crucial role in helping me iron out details regarding this project. I also received help from our TA, Mr. Soumyajyoti Bannerji in resolving how to encode the categorical values. Thank you.